ARC CENTRE OF EXCELLENCE FOR
THE DYNAMICS OF LANGUAGE

Referencing Data in Publications: principles, policy and practice

Linguistics

Julia Colleen Miller

julia.miller@anu.edu.au

28 October 2015, Jaeger Room, The Shine Dome

The University Of Queensland AUSTRALIA

WESTERN SYDNEY UNIVERSITY

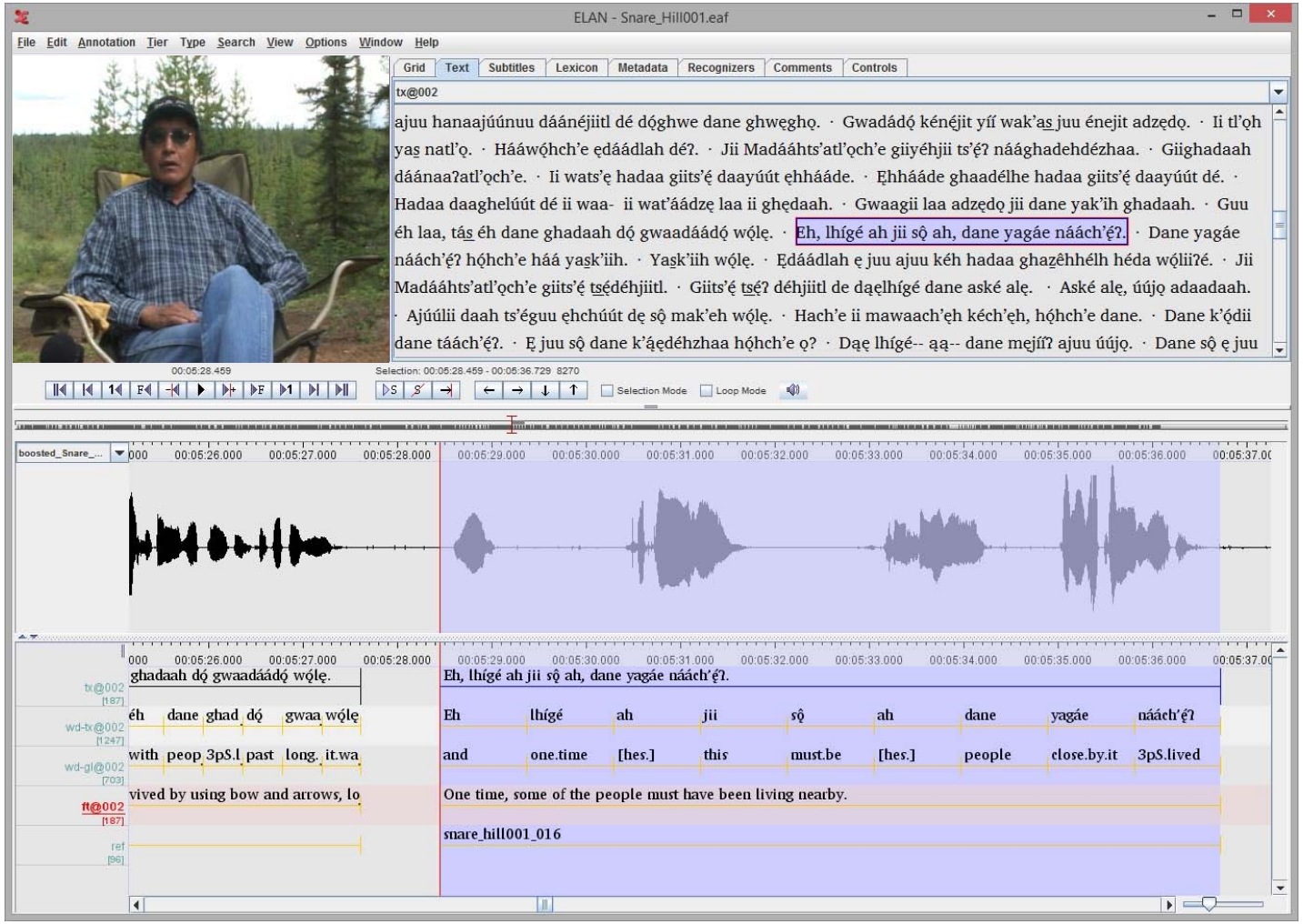
Australian Government
Australian Research Council

Australian National University

THE UNIVERSITY OF MELBOURNE

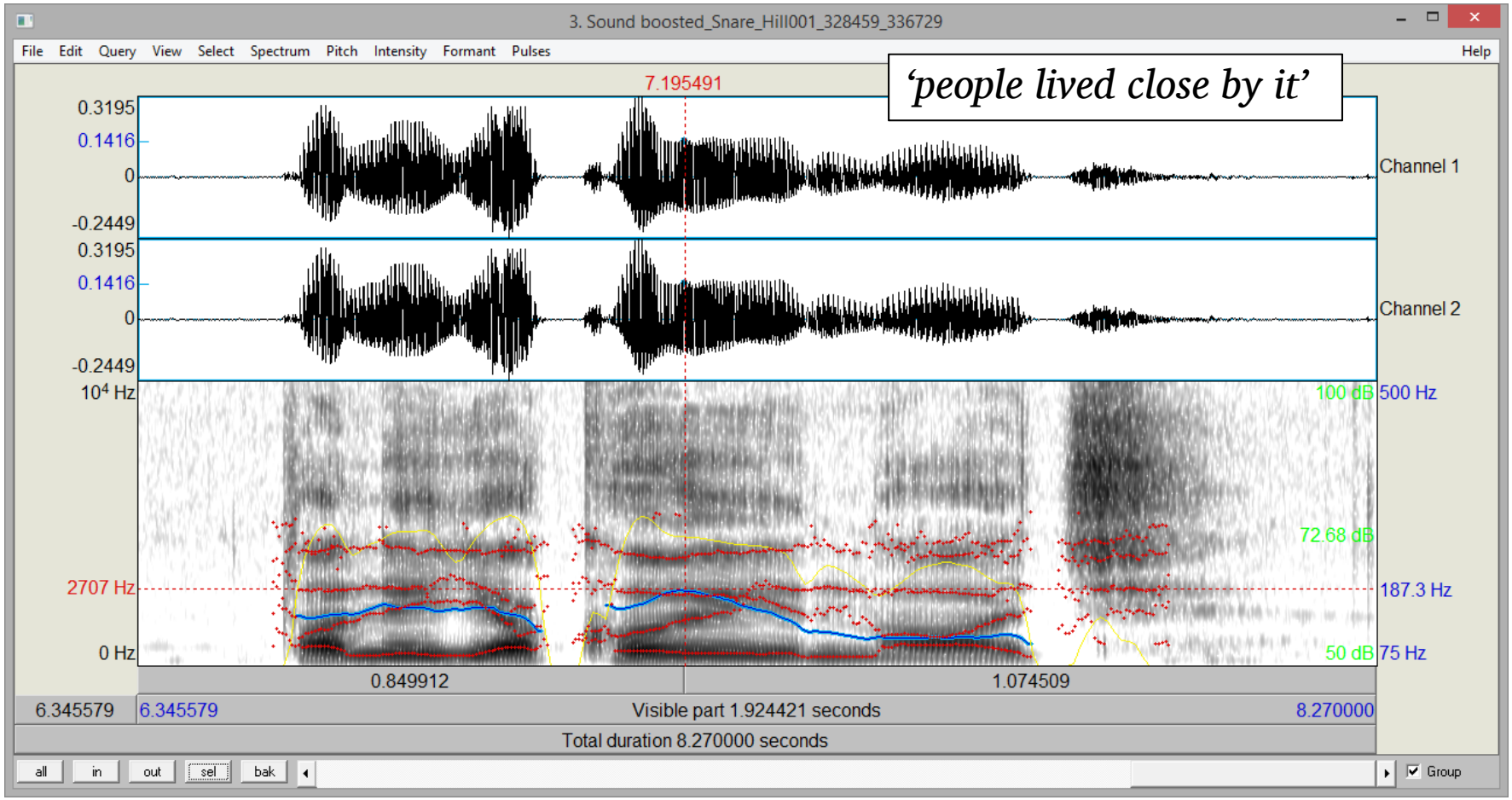
Shape of linguistic data

ARC CENTRE OF EXCELLENCE FOR THE DYNAMICS OF LANGUAGE



Shape of linguistic data

ARC CENTRE OF EXCELLENCE FOR
THE DYNAMICS OF LANGUAGE



Shape of linguistic data

ARC CENTRE OF EXCELLENCE FOR THE DYNAMICS OF LANGUAGE

Granularity: a corpus (collection)

PARADISEC

PARADISEC Catalog

Julia Colleen Miller | Sign out

Home | Dashboard | Collections | Items | Contact

Return To Results

Edit collection

Collection details

Ready for metadata export Private: Hide metadata from all users

Collection ID	SocCog
Title	Social Cognition Project
Description	Materials from a structured task for gathering enriched language data for descriptive, comparative and documentary purposes. The task involves collaborative narrative problem-solving and retelling by a pair or small group of language speakers, and was developed as an aid to investigating grammatical categories relevant to psychosocial cognition. The pictures set up a dramatic story where participants can feel empathetic involvement with the characters, and trace individual motivations, mental and physical states, and points of view. The data-gathering task allows different cultural groups to imbue the pictures with their own experiences, concerns, and conventions and stimulates the spontaneous use of previously under-recorded linguistic structures. Picture sets include "Family Problems" and "Crow and Jackal".
Archive link	http://catalog.paradisec.org.au/repository/SocCog
Collector	Nicholas Evans
Operator	Julia Colleen Miller
Originating university	Australian National University
Countries	
Languages	
Region / village	Various

Items in Collection (17)

Add new item

Item	Title	Digitised	Files	Actions
001 open	Elicitation materials - "Family Problems"		17	View Edit
002 open	Elicitation materials - "Crow and Jackal"		9	View Edit
101 open	Family Problems in Momu recorded in Mumuru Village		10	View Edit
102 open	Family Problems in Burmese		2	View Edit
103 open	Family Problems Yurakaré Alex and Mercy		17	View Edit
104 open	Family Problems Yurakaré Gerónimo and Julia		18	View Edit
105 open	Family Problems Yurakaré Elizabet and Gladis		18	View Edit
106 open	Family Problems Yurakaré Benancio and Celia		18	View Edit
107 open	Family Problems Yurakaré Arsenio and Gladis		18	View Edit
108 open	Family Problems Yurakaré Pablo and Rebeca		18	View Edit
109 open	Family Problems Yurakaré Jedeón and Juan Carlos		18	View Edit
FLEX open	Sanzhi Flex		1	View Edit
HMK open	Sanzhi Social Cognition Elicitation		14	View Edit
Iwaidja open	Family Problems Stories: Iwaidja		0	View Edit
MZ open	Sanzhi Social Cognition Elicitation		10	View Edit
PX open	Sanzhi Social Cognition Elicitation		9	View Edit
TA open	Sanzhi Social Cognition Elicitation		9	View Edit

Show 10 | Show 50 | Show all

Shape of linguistic data

ARC CENTRE OF EXCELLENCE FOR THE DYNAMICS OF LANGUAGE

Granularity: a text within a corpus (item)

PARADISEC

PARADISEC Catalog

Julia Colleen Miller | Sign out

Home | Dashboard | Collections | Items | Contact

Return To Results | Previous item | Edit item | Duplicate Item | Next item

Item details

Private: Hide metadata from all users

Item ID	SocCog-MZ (Collection Details)
Title	Sanzhi Social Cognition Elicitation
Description	Social cognition picture task. 001 files offer picture descriptions; 002 files offer picture ordering; 004 files offer story telling (Muslimat). There is one video file belonging to this and two audio files. Speakers: Muslimat, Zhavhar.
Origination date	2014-02-03
Origination date free form	
Archive link	http://catalog.paradisec.org.au/repository/SocCog/MZ
URL	
Collector	Diana Forker
Countries	
Language as given	Sanzhi Dargwa
Subject language(s)	
Content language(s)	

Content Files (10) View all

Filename ▲▼	Type ▲▼	File size ▲▼	Duration ▲▼	File access
SocCog-MZ-Sanzhi_03_02_2014_DF_003_short.eaf	application/xml	50.1 KB		View
SocCog-MZ-Sanzhi_03_02_2014_DF_001.eaf	application/xml	126 KB		View
SocCog-MZ-Sanzhi_03_02_2014_DF_003.wav	audio/x-wav	406 MB	00:24:38.190	View
SocCog-MZ-Sanzhi_03_02_2014_DF_003.mp3	audio/mpeg	22.5 MB	00:24:38.240	View
SocCog-MZ-Sanzhi_03_02_2014_DF_003_short.wav	audio/x-wav	81.3 MB	00:07:22.632	View
SocCog-MZ-Sanzhi_03_02_2014_DF_003_short.mp3	audio/mpeg	6.74 MB	00:07:22.670	View
SocCog-MZ-Sanzhi_03_02_2014_DF_001.wav	audio/x-wav	350 MB	00:21:12.279	View
SocCog-MZ-Sanzhi_03_02_2014_DF_001.mp3	audio/mpeg	19.4 MB	00:21:12.319	View
SocCog-MZ-Sanzhi_03_02_2014_DF_002.wav	audio/x-wav	197 MB	00:11:55.605	View
SocCog-MZ-Sanzhi_03_02_2014_DF_002.mp3	audio/mpeg	10.9 MB	00:11:55.650	View
	--	**1.07 GB**	--	--

Show all

Cite as Diana Forker (collector), 2014; *Sanzhi Social Cognition Elicitation*, XML/X-WAV/MPEG, http://catalog.paradisec.org.au/collections/SocCog/items/MZ 2015-10-27.

Shape of linguistic data

ARC CENTRE OF EXCELLENCE FOR THE DYNAMICS OF LANGUAGE

Granularity: a timecoded selection of the text

Shape of linguistic data

ARC CENTRE OF EXCELLENCE FOR THE DYNAMICS OF LANGUAGE

Information about the text (item)

Collection: SocCog

Item: MZ

Contributors:

- Diana Forker (compiler)

Description:
Social cognition picture task. 001 files offer picture descriptions; 002 files offer picture ordering; 004 files offer story telling (Muslimat). There is one video file belonging to this and two audio files. Speakers: Muslimat, Zhavhar. . Language as given: Sanzhi Dargwa

Cite this item:
Diana Forker (collector), 2014; Sanzhi Social Cognition Elicitation, XML/X-WAV/MPEG, http://catalog.paradisec.org.au/collections/SocCog/items/MZ 2015-10-27.

Zotero and Mendeley Integration.

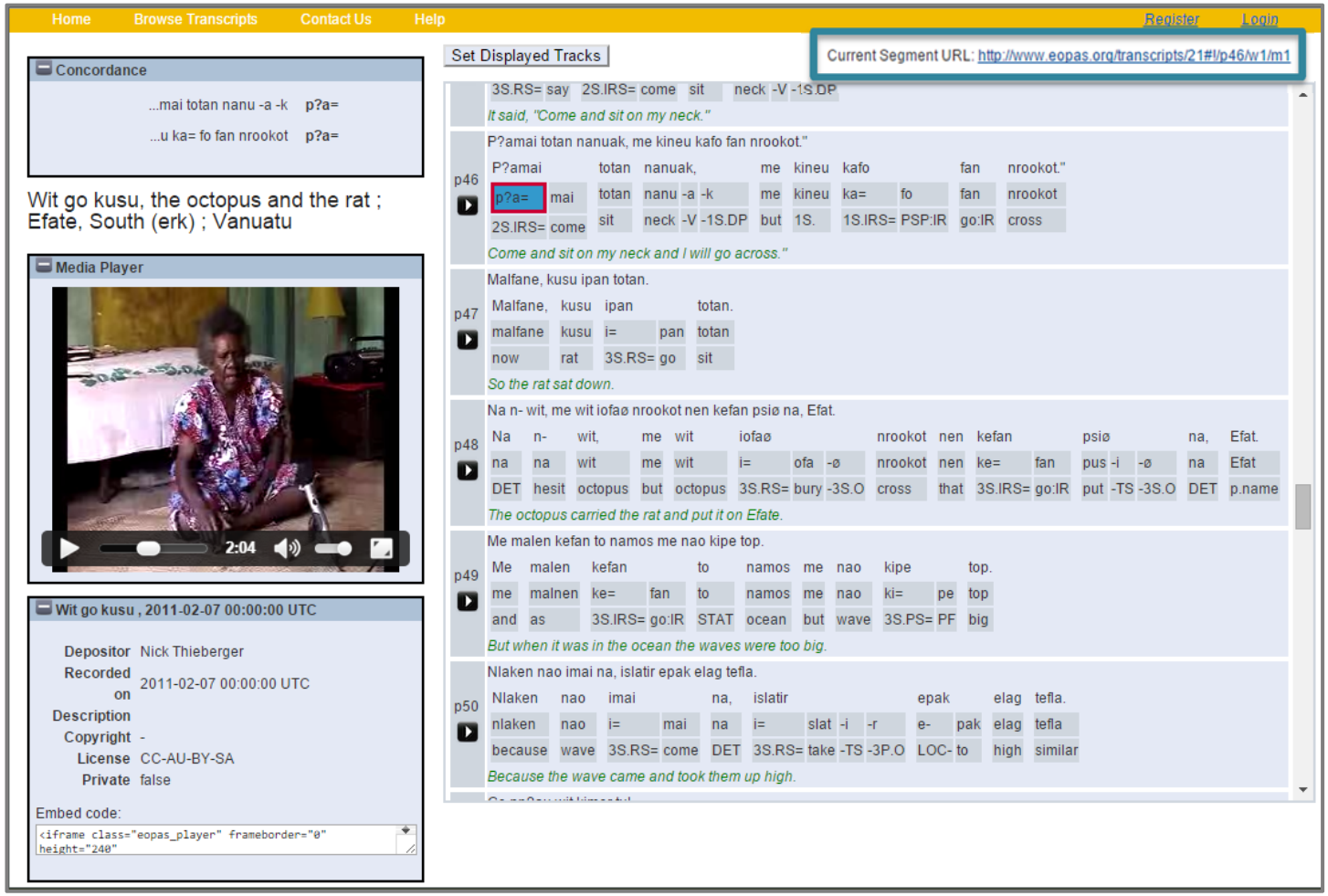
This page has embedded metadata that is able to be read by both Zotero and Mendeley.

Zotero users with the appropriate browser plugin (https://www.zotero.org/download/) can save to Zotero directly whilst Mendeley users should install the *one click Web Importer* (https://www.mendeley.com/import/) to enable the same functionality.

Shape of linguistic data

ARC CENTRE OF EXCELLENCE FOR THE DYNAMICS OF LANGUAGE

Granularity: a single morpheme within the timecode



How is data referenced?

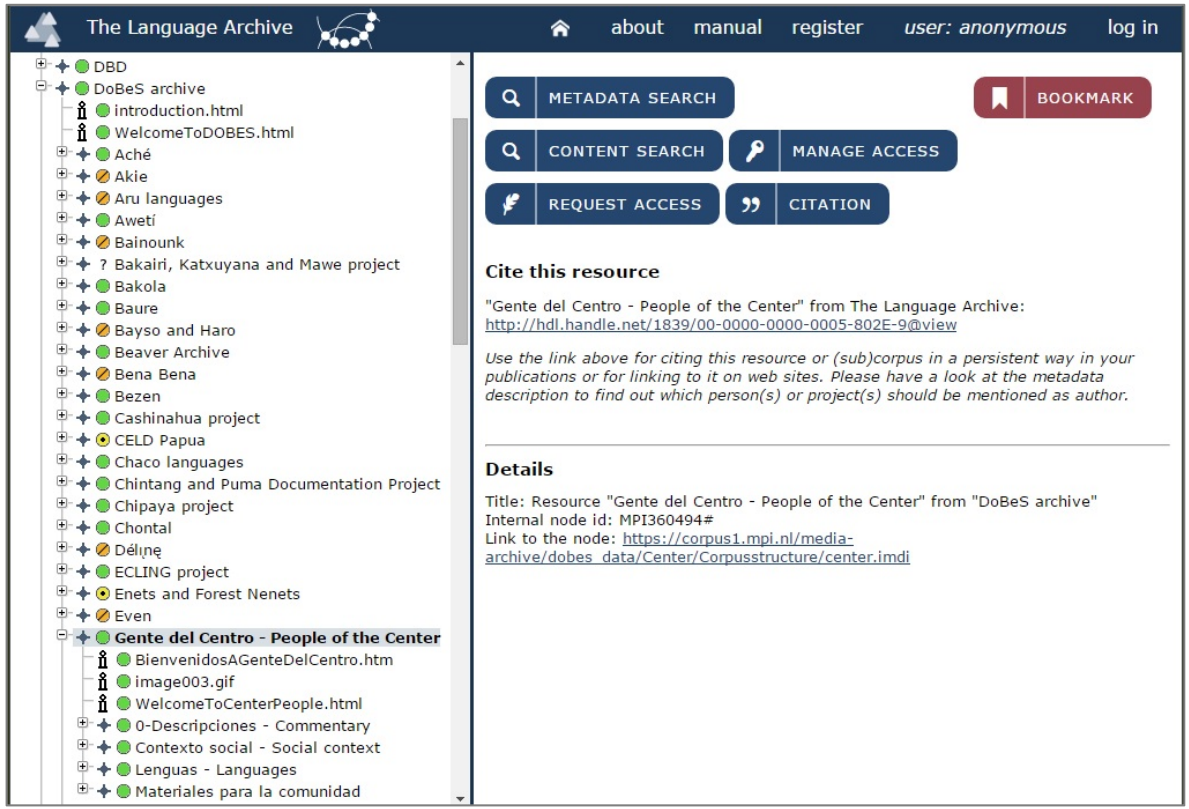
ARC CENTRE OF EXCELLENCE FOR THE DYNAMICS OF LANGUAGE

Example of a corpus referenced in the article *Direct and indirect affix borrowing (2015),* published in the journal of the Linguistics Society of America, Language.

SEIFART, FRANK. 2009. Resígaro documentation. *A multimedia documentation of the languages of the People of the Center: Online publication of transcribed and translated Bora, Ocaina, Nonuya, Resígaro, and Witoto audio and video recordings with linguistic and ethnographic annotations and descriptions*, ed. by Frank Seifart, Doris Fagua, Jürg Gasché, and Juan Alvaro Echeverri. Nijmegen: DOBES-TLA. Online: http://corpus1.mpi.nl/qfs1/media-archive/dobes_data/Center/Info/WelcomeToCenterPeople.html.

How is data referenced?

ARC CENTRE OF EXCELLENCE FOR THE DYNAMICS OF LANGUAGE



The guide outlines:

- who was involved in the documentation and what kind of materials were collected and why
- how to find data in the archive
- access rights
- citation rules
- authorship & acknowledgements
- suggestions as to where to begin searching the corpus

DOBES-archive for the languages of the People of the Center, North West Amazon: An Introduction

(pasar a versión española)

Frank Seifart, Doris Fagua, Jürg Gasché, and Juan Alvaro Echeverri

July 2009, updated May 2010

Welcome to the DoBeS documentation of the languages Bora, Huitoto, Nonuya, Ocaina, and Resígaro, which belong to a group of people known as People of the Center (orChildren of the Tobacco, Coca and Sweet Manioc). This documentation contains about 2,000 audio and video recordings, many of them transcribed and translated, as well as additional linguistic and ethnographic descriptions. The present document is a guide for anyone interested in consulting this documentation. It contains important information about how the archive is structured and how the data and descriptive documents can be accessed. It is highly recommended to read this document before entering the archive itself.

The DoBeS People of the Center Documentation project

The People of the Center are a multilingual group living in the North West Amazon, in the border region between Southern Colombia and Northern Peru. Seven mutually unintelligible languages belong to the People of the Center: Andoke, Bora, Huitoto, Muinane, Nonuya, Ocaina, and Resígaro. The People of the Center are characterized by some unique cultural practices, including elaborate ritual discourses, festivals at which repertoires of thousands of songs are performed, as well as systems of drum communication that build on the structures of the individual languages. The traditional cultural practices as well as the use of the languages on a daily basis are nowadays rapidly declining in a process of transculturalization towards national society and language shift towards local Spanish. Since 2004 the Volkswagen Foundation supports the project "Documenting the languages of the People of the Center", which was housed from 2004-2007 at the Ruhr-Universität Bochum and from 2007-2009 at theUniversität Regensburg. The partner organization in Peru is the Instituto de Investigaciones de la Amazonia Peruana (IIAP). The project focuses on the languages Bora, Huitoto, Nonuya, Ocaina, and Resígaro, documenting the use of each of these languages in different contexts, ranging from memorized ritual discourses to traditional festivals, hunting and fishing trips, to sport events and everyday conversation. The aim of this documentation is to provide a lasting record of the cultural and linguistic practices of People of the Center as they can be observed at the beginning of the 21st century. In order to ensure that members of the People of the Center themselves can access and make use of this documentation, all translations, descriptions, and metadata are written in Spanish. Copies of the entire documentation are stored in two physical locations, the central DOBES archive in Nijmegen and the regional archive in Iquitos.

How to find the data and descriptions in the archive

The archive is divided into two main components: A corpus of primary data and a commentary, which includes a set of descriptive and introductory documents. In the corpus of primary data, the audio and video recordings are stored together with "annotations", i.e. transcriptions and translations, in some cases also morphological segmentation and glossing. Each "bundle" consisting of a recording and an annotation is called a "session". There are basically two ways to find a particular session in the corpus. First, you can explore the corpus by clicking on the nodes on the left. This will gradually unfold the structure of the archive (a short introduction can be found here). When reaching a session node, you may obtain some important information about the corresponding recordings and annotations by clicking on the session-"bundle" symbol (). This information is called the "metadata".

The sessions will be organized into two parallel corpus trees: according to languages and according to social context. At the moment (as of March 2010), sessions are only linked in the corpus tree, which organizes the sessions according to languages, which includes sub-nodes corresponding to broad categories such as "Songs", "Conversation", etc. A second way to find particular sessions is by searching the metadata of corpus of the People of the Center. Right-click on the node "Gente del Centro / People of the Center", select "metadata search" and enter the keywords you want to search for.

The second main component of the archive contains a set of descriptive and introductory documents, such as grammar sketches, introductory text to the history, social organization, etc. of the People of the Center. These

documents may be accessed by browsing under "Commentary". All of these documents are open access. It should be noted that many of these are still preliminary versions (as of March 2010).

Access rights, quoting conventions

The access to the metadata (the information about each session) is open to anybody. However, the access to much of the data itself (i.e. recordings and annotations) is restricted, i.e. you may have to ask for permission to view or download them (see below). The main reasons for this are privacy rights of the speaker communities and the need to maintain some kind of control over the use of the data (see also http://www.mpi.nl/DOBES/ethical_legal_aspects/). If you are interested in viewing or downloading a recording or annotation with restricted access, please do not hesitate to request access to it, through right-clicking on the file-symbol (*.wav, *.mpg, etc. for recordings; *.txt, *.eaf, etc. for annotations) and selecting "request resource access" (alternatively, send an email to frank.seifart_at_berlin.de). Every request to access data will be considered by us.

When quoting from or referring to data contained in this documentation, please use the following reference:

Seifart, Frank, Doris Fagua, Jürg Gasché, and Juan Alvaro Echeverri (eds.). 2009. *A multimedia documentation of the languages of the People of the Center. Online publication of transcribed and translated Bora, Ocaina, Nonuya, Resígaro, and Witoto audio and video recordings with linguistic and ethnographic annotations and descriptions*. Nijmegen: DOBES-MPI.

When quoting only data from one language, you may refer to the sub-corpus and the person responsible for compiling it (see next section), for example:

Fagua, Doris. 2009. Ocaina Documentation. In Seifart, Frank, Doris Fagua, Jürg Gasché, and Juan Alvaro Echeverri (eds.). *A multimedia documentation of the languages of the People of the Center. Online publication of transcribed and translated Bora, Ocaina, Nonuya, Resígaro, and Witoto audio and video recordings with linguistic and ethnographic annotations and descriptions.* Nijmegen: DOBES-MPI.

To quote individual nodes or sessions of the corpus, a reference can also be obtained by right-clicking the node or session symbol and choosing "create bookmark".

When quoting from or referring to the descriptive and analytical documents stored here, please refer to the author(s) and the title of the document and make use of the reference date given at the beginning of each document. Thus, when quoting from the Witoto (Huitoto) grammar sketch deposited here, give the following reference as your source:

Gasché, Jürg. 2009. Esbozo gramatical de la lengua huitoto. In Seifart, Frank, Doris Fagua, Jürg Gasché, and Juan Alvaro Echeverri (eds.). *A multimedia documentation of the languages of the People of the Center. Online publication of transcribed and translated Bora, Ocaina, Nonuya, Resígaro, and Witoto audio and video recordings with linguistic and ethnographic annotations and descriptions.* Nijmegen: DOBES-MPI.

Authorship, editorship, acknowledgements

The authors of the primary data deposited in this archive are the speakers of the language of the People of the Center, as specified for each session in the accompanying metadata. The sub-corpora on individual languages have been compiled under the responsibility of the project members specializing on these languages. These are: Bora: Frank Seifart and Jürg Gasché; Witoto (Huitoto): Jürg Gasché; Nonuya: Juan Alvaro Echeverri; Ocaina: Doris Fagua; Resígaro: Frank Seifart.

The speakers of the languages of the People of the Center not only agreed to deposit their linguistic and cultural heritage in this archive, they also supported this project in many other ways, including transcription and translation and logistic support. We are grateful to them as well as to the regional indigenous organization FECONA for their institutional support.

We would like to thank Nikolaus Himmelmann for initially applying for this project with us, for housing it at his department at the Ruhr Universität Bochum, and for invaluable academic as well as administrative support. We are also grateful for support of various kinds from the following people in Bochum: Martin Hölter, Jan Strunk, Victoria Rodriguez, Judith Köhne, Rahel Beyer, and Iva Renic.

We would like to also thank the student assistants at the Universität Regensburg: Katharina Höhendinger, Melissa Houser, Dominik Messer, Susanne Naimer, Katharina Knuffmann, Nicole Goltz, Vanessa Messer, Florian Greiner, and Michael Fischbach.

In Iquitos, project work was supported by Ernesto Tello, Zacarias Mibeco, Nandier Meza, and Joilé Villanueva, and in Bogotá by Edilamar Bustos.

En el IIAP, agradezemos la por la fructífera cooperación al presidente Dr. Luis Campos Baca, al responsable de la cooperación técnica Econ. Hernán Tello y al responsable del servicio informático, Ing. Victor Miyakawa.

Suggestions where to start

The following are some examples of interesting sessions that are accessible without restrictions, and that you may consult to get a first impression about the People of the Center. In order to view these, open the session node ( ) and right-click on the transcription (*.eaf) file (if there is one) or the video (*.mpg) file:

Bora:	The story of the daughter of the creator: Piivyeebe_ajyu Fishing with poison: wakyu_cant_1
Huitoto:	Songs and conversation among men at night in the "coca square": Lluaky1
Ocaina:	Making a mousetrap: Jopahtya1 and Jopahtya11 A presentation of songs and dance at a festival for a new house: Juhtoxohxo2
Resígaro:	Some words of one of the last mother-tongue speakers of Resígaro: storyalicia

Once you see the hierarchical corpus structure, we recommend selecting under "settings" (top right corner) "Show accessible resources". This will display in square brackets after each corpus node the number of accessible resources for the current user (by default "anonymous") and therefore may help you to locate additional resources that are currently available without registration.

DVD and CD library

A collection of DVDs with bilingual subtitles and audio CDs was created from the recordings, transcriptions, and translations made in the context of this project, as part of thecommunity materials.

Related documentations

Data from Andoke, a language of the People of the Center which we were not able to document within the present project, were collected by Jon Landaburu and are archived in the Archive of the Indigenous language of the Americas (AILLA) (see http://www.ailla.utexas.org/search/view_resource.html?lg_id=103). Data from Miraña, a close linguistic variant of Bora, spoken by the Miraña people in Colombia, are deposited here: http://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI33493%23.

How is data referenced?

ARC CENTRE OF EXCELLENCE FOR THE DYNAMICS OF LANGUAGE

LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 8 (2014), pp. 490-524
http://nflrc.hawaii.edu/ldc/
http://hdl.handle.net/10125/24615

Finding a way into a family of tone languages: The story and methods of the Chatino Language Documentation Project

Emiliana Cruz
University of Massachusetts at Amherst

Anthony C. Woodbury
University of Texas at Austin

(3)	a.		ʔo	+	yja	-->	ʔo	yja	'and tortillas'
			with		tortilla				
		Phonemic:	/H/		-				
		Phonetic:	[H]		[L]		[H]	[ML]	
	b.		ʔo	+	nda	-->	ʔo	nda	'and beans'
			with		bean				
		Phonemic:	/H/		/L/				
		Phonetic:	[H]		[L]		[H]	[L]	

Click to activate...

Moving forward

ARC CENTRE OF EXCELLENCE FOR THE DYNAMICS OF LANGUAGE

Journals:

- Primary data should be provided with sources
- Persistence of the source is important
- Create a unified style-sheet for citing data

Archives:

Annotated corpora counted as a referred publication is a foreseeable vision for the future.

- criteria for reviewing a corpus
- provide, clear instructions for citing the archived data

Moving forward

ARC CENTRE OF EXCELLENCE FOR THE DYNAMICS OF LANGUAGE

Linguistic data:

- should be seen as legitimate, citable research
- should allow for data contributors to be given proper attribution
- whenever there is a claim, data should be provided as evidence
- there should be a unique identifier associated (DOIs) with the data. granularity to be agreed upon by peak bodies; this identifier should be persistent
- access to data and associated material should be facilitated by citation, and data should be accessible without the use of special software or hardware.

Thank you!

PARADISEC Catalog ***http://catalog.paradisec.org.au/***

https://corpus1.mpi.nl/ds/asv/?1

ARC CENTRE OF EXCELLENCE FOR
THE DYNAMICS OF LANGUAGE ***http://www.dynamicsoflanguage.edu.au/***