



Investigation of descriptive richness of free-text metadata in language archives

Mary Burke, Oksana L. Zavalina
University of North Texas
mary.burke@unt.edu,
oksana.zavalina@unt.edu



Overview

- Introduction: what is a language archive?
- Methodology
 - Research questions
 - Data collection
 - Data analysis
- Results & Examples
- Discussion & what's next?



Digital Language Archives



- ✓ 2+ items
- ✓ structured metadata
- ✓ open to deposits
- ✓ goals: preservation, access

- not necessary to self-identify as a language archive



Introduction

- Language archives
 - Managed/created by linguists in 1990's
 - Use locally-developed schemes
 - Struggle with authority control, ease of data re-use (Burke & Zavalina, 2019; 2020)
- Metadata quality highly variable
 - Content management systems/ archival software
 - Self-upload v. mediated deposit
 - Metadata creation guidelines

Identifier:	ANLC3234
Title	SLI De Reuse #3
Description:	Grammatical comments on text on singing. The morning of 4/12. Slow repeat of 2nd episode of 4/12
Comments	#3 WdR 3a+b. Tape is located in Annex. Tape 3 of 35.
Contributors	de Reuse, Willem Joseph (interviewer)
Date	1985-04-05
Type	Sound
Subject Language(s)	Siberian Yupik
Collection	SLI De Reuse

Research Questions & Method

Manual content analysis

What are the **differences and similarities** across the 3 archives:

- in the types of information included in free-text Description fields of metadata records?
- in quantitative characteristics of the Description field data values: **lengths of data values, the number of categories of information included, etc.?**

Data Collection

- 3 language archives
 - Endangered Languages Archive (**ELAR**)
 - Pacific Regional Archive for Digital Sources in Endangered Cultures (**PARADISEC**)
 - Archive of the Indigenous Languages of Latin America (**AILLA**)
- Random selection of 130 item-level metadata records from each
- Data values of the Description metadata fields ONLY:
 - 390 instances of Description field data values in English language exported to an Excel spreadsheet



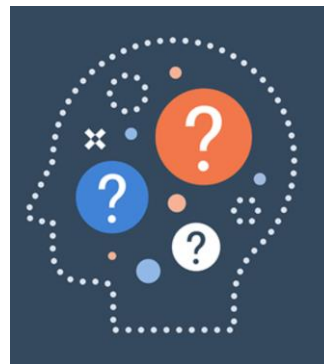
Descriptions

The argumentative stimulus contain argumentative situations, which were read by one of the participants in recordings to other participants. After the situations were read, the participants engage themselves in argumentation on the read situations. The items

Family Problems Yurakaré Alex and Mercy

160506-000 ; About the earthquake ; Karsang talks about the earthquakes of 2015 in Nepal ; Speaker(s): Karsang ; Audience: Dawa Chenzom ; Relationship: spouses ; Location: Dawas kitchen KyanjGompa ; Genre: Report - explanatory ; BOLD careful respeaking and separate file with Nepali translation. | 160506-001 ; About the earthquake and rebuilding ; Ngawang talks about the earthquake of April 2015 and rebuilding, after being asked 'how will you build your new house?' ; Speaker(s):Ngawang ; Audience: Karsang ; Relationship: friends ; Location: Dawas kitchen KyanjGompa ; Genre: Report - explanatory ; BOLD careful respeaking and separate file with Nepali translation. | 160506-002 ; How will you rebuild ; Ngawang and Karsang talk about the earthquake of April 2015 and rebuilding, after being asked 'how will you build your new house?'. Can cut ending of recording ; Speaker(s):Ngawang, Karsang ; Audience: None ; Relationship: friends ; Location: Dawas kitchen KyanjGompa ; Genre: Report - explanatory ; BOLD careful respeaking and separate


Data Analysis



- Coding categories based on:
 - types of information observed in descriptions
 - metadata creation guidelines of language archives
 - related previous study of Description field data in collection-level metadata records (*Zavalina et al., 2008*)

→ **28 total categories (shown on the next slide)**

- Binary coding, 0 or 1
- Validation: 20% (78 descriptions) coded by 2 coders
 - 95+% agreement
- Descriptive statistics: word count, number of categories

Genre	Description template	Example
		
Retelling	This is a retelling of { } narrated by { }. The story is about {No more than 100- word description of story}.	This is a retelling of the Pear Story narrated by Beshot Khullar. In this story a boy steals a basket of pears and meets and shares pears with three other boys.

28 Coding Categories



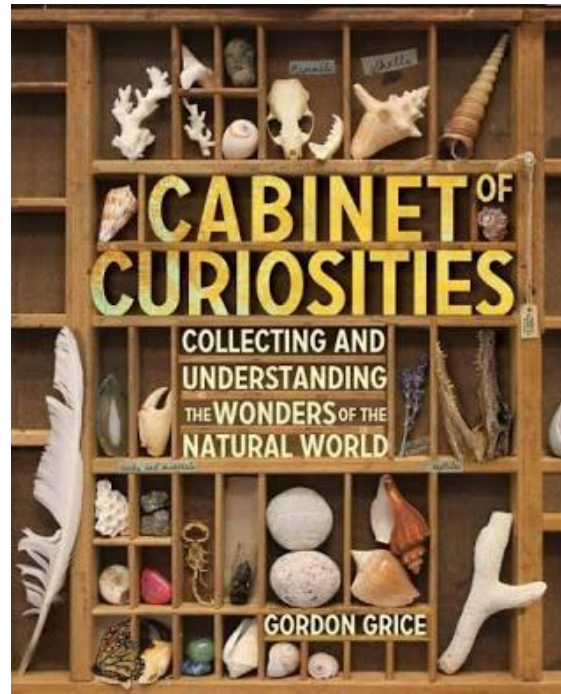
Summary of content	Names/ roles of speaker(s)	Audience/ uses	Provenance
Genre	Recording equipment / technical issues	Grammatical constructions	Notes from tape/ box
Transcription or translation of recording	Sociolinguistic information	Bibliographic citation	Partner institution/ funding agency
Translation of description	Related items	Administrative notes	Languages/ dialect
Geographical coverage	Access rights	Linguistic/ cultural context	Stimuli used
Temporal coverage	Value/ uniqueness	File names	Date of creation
Creator/ Depositor/ Translator	Quality	Setting/ context of recording	Title



Selected examples of Description field data values



- Story about POP's life
- Side A: Bilua Words 1-134; Side B: 2 texts and words 135-209
- Ungwa:Cane
- Greenland Eskimo
- Fauna
- All audio recordings created using Zoom H4n audio recorder with internal mic (.wav format 44.1 kHz, 16-bit stereo).



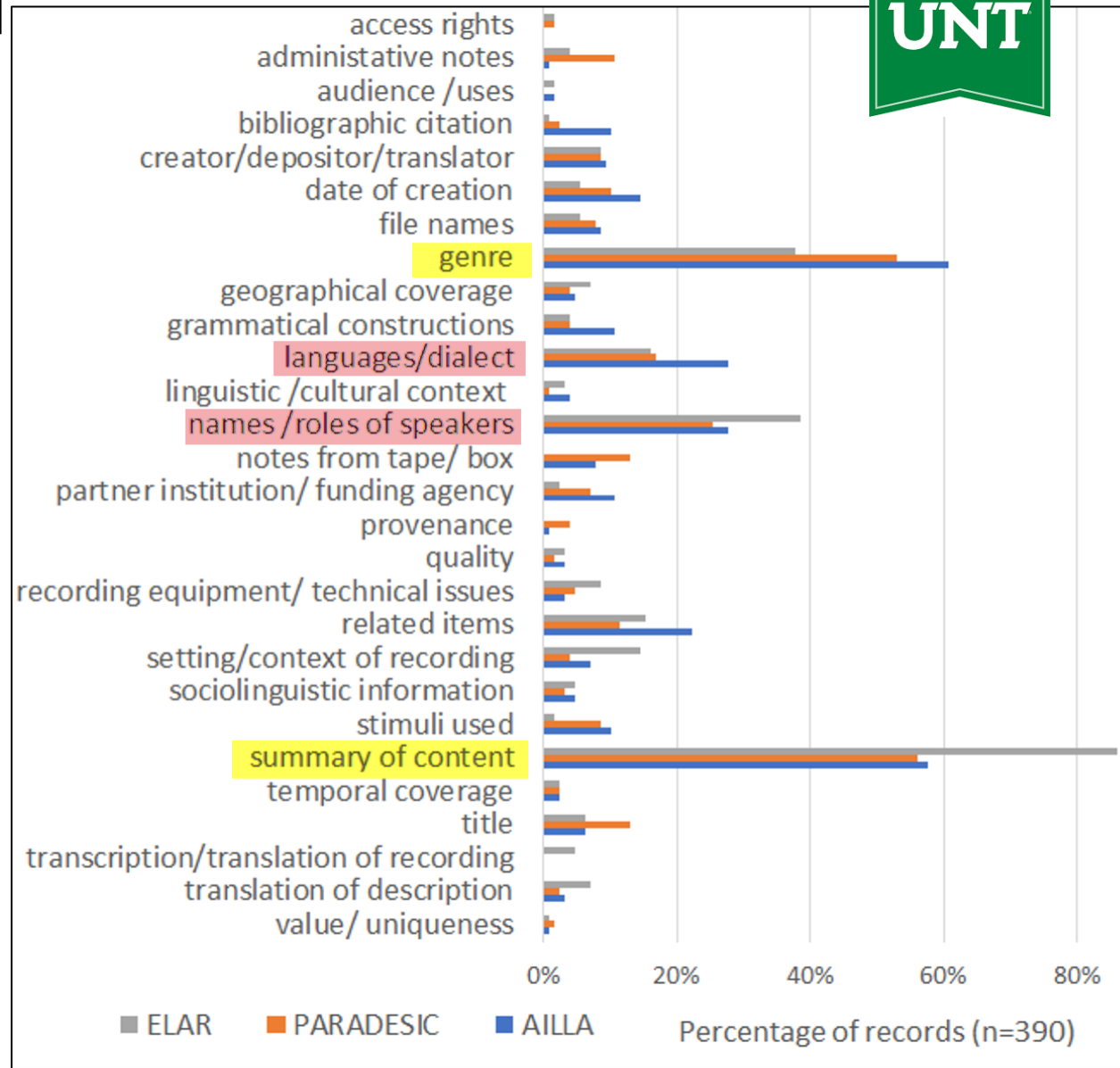
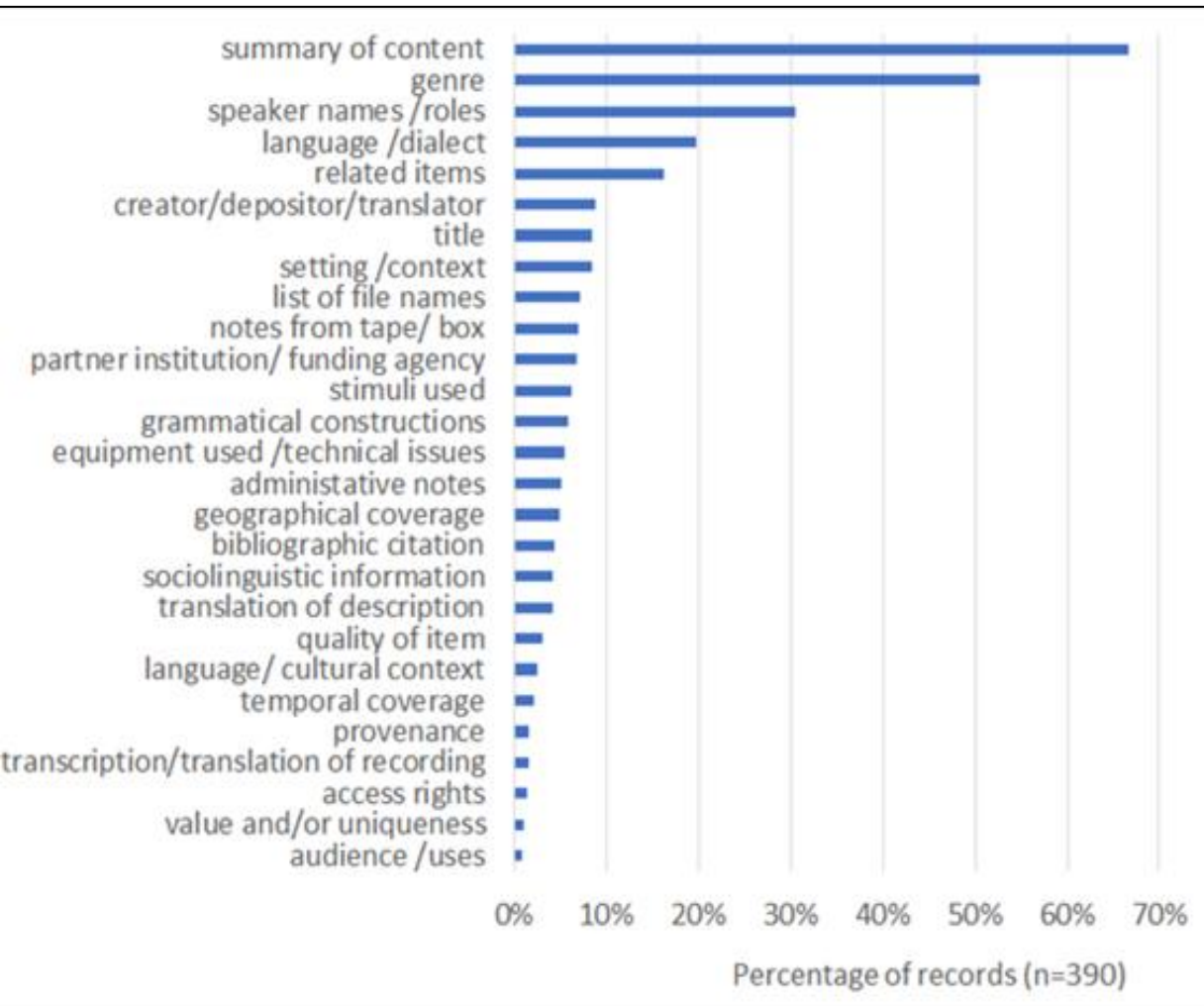
- ZF1-Kikamba_9_11_90-A elicitation
- ZF1-Kikamba_9_11_90-B elicitation
- Tapsut (Dahaplan); Tilit (Men's house); Henna (MPI reciprocals)
- To cure a stomach ache - (malikái)
- A/R = Rabi
- An anthill

Results: descriptive statistics

Indicator:	Minimum	Maximum	Average	Median	Standard Deviation
length of data value in words	1	706	52.9	20.5	83.1
number of description categories observed	0	10	2.8	2	1.7



Results: Categories of information observed



Discussion

- Metadata creators = linguists
 - insufficient training/ institutional support
 - Better understanding of language archive user needs
- develop metadata creation guidelines
- examples in AILLA guidelines (AILLA, n.d.)

“not only the description of the session **but also information about its content**” (ELAR, n.d.)

Description [English]/
Descripción [inglés]

Description of the resource in English/
Descripción del recurso en inglés

Adolfo Santiago tells a tale of the twins the Sun and the Moon.

“Four text stories for interviews”
(PARADISEC, n.d.)

What's next?



- **Larger scale comparison of metadata quality in language archives**
 - Include other language archives (e.g., [DELAMAN](#) archives)
 - Create mapping of common metadata elements used in language archives
 - Analyze quality of entire metadata records (beyond Description field)



Works Cited

- Archive for the Indigenous Languages of Latin America (n.d.). *Metadata*. Accessed 1 September 2020.
- Burke, M., Zavalina, O. L., (2019). Exploration of Information Organization in Language Archives. *Proceedings of the Association for Information Science and Technology* 56, 364-367.
- Burke, M., Zavalina, O. L., (2020). Identifying challenges for information organization in language archives: Preliminary findings. *Proceedings of iConference 2020*.
- Endangered Languages Archive (n.d.). *Depositing with ELAR*. Accessed 1 September 2020.
- Pacific Regional Archive for Digital Sources in Endangered Cultures. (n.d.) *Depositing material with PARADISEC*. Accessed 1 September 2020.
- Zavalina, O. L. Palmer, C. L., Jackson, A. S., and Han, M. (2008). Evaluating descriptive richness in collection-level metadata. *Journal of Library Metadata* 8(4): 263–292.





Questions?
Suggestions?
Possible collaborations?

mary.burke@unt.edu
oksana.zavalina@unt.edu

