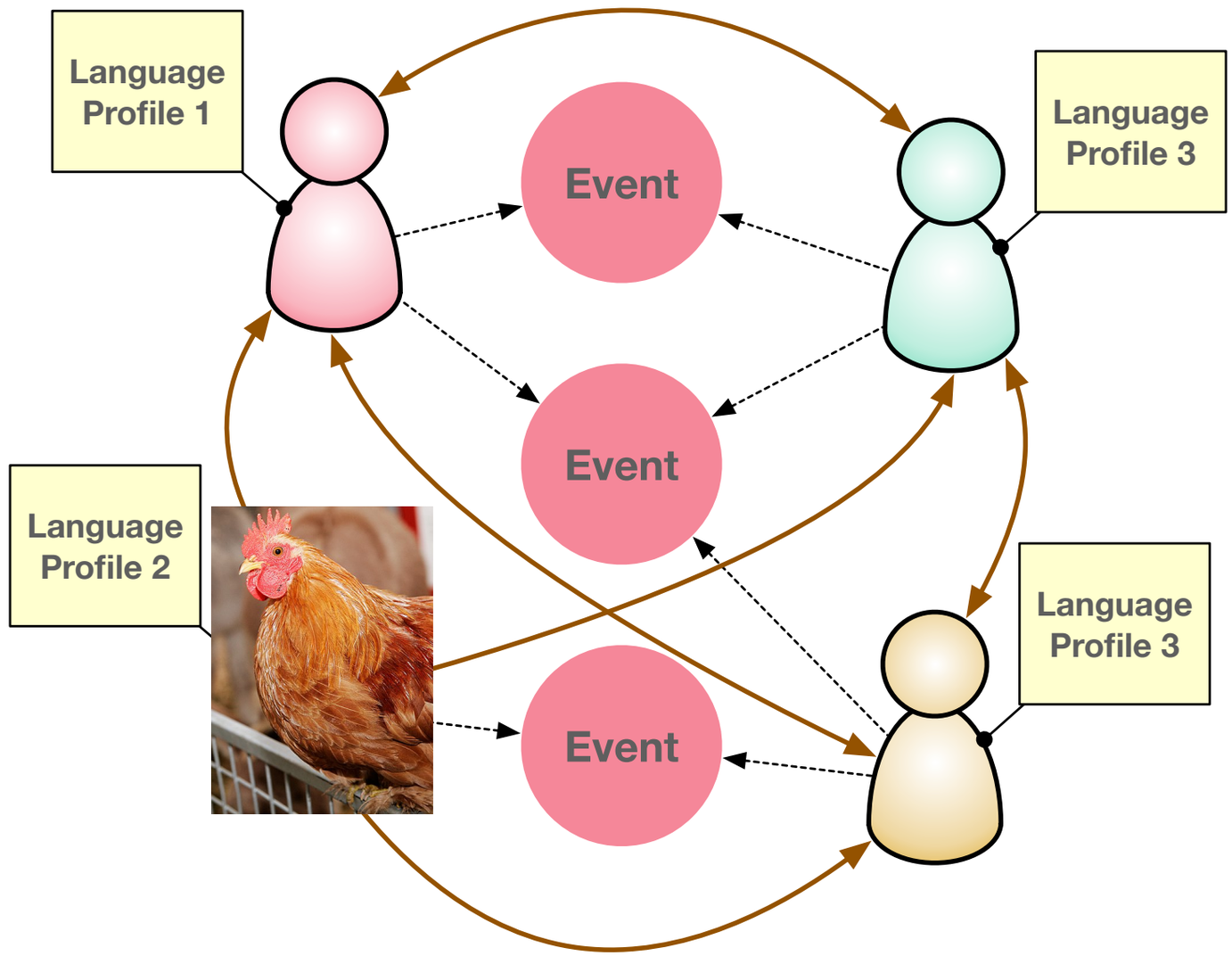
Extending CLDF to Multilingual (Text) Data

Source: Fir0002/Flagstaffotos

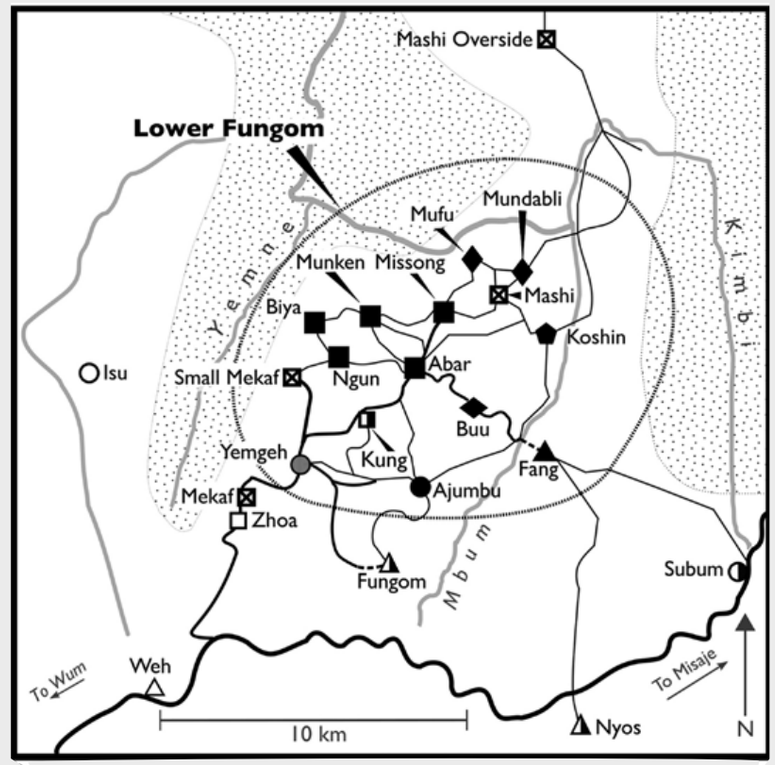


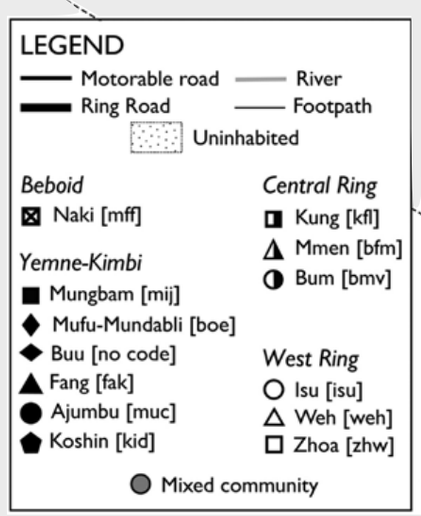
Jeff Good (jcgood@buffalo.edu)
University at Buffalo and Humboldt-Universität zu Berlin

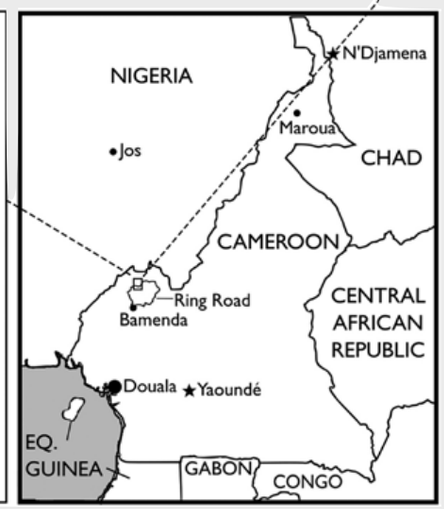
14–15 December 2023, MPI-EVA
Cross-Linguistic Data Formats Workshop

An opening multilingual exchange

Lower Fungom, Cameroon







For more information about the project this work is based on, visit the KPAAM-CAM website at: https://kpaam-cam.org/.

Kulani to Pa Joshua

Pa Jo! iyə!

Pa Jo! How are you?

aboh koh daŋ koh oti corner mohngən noh

You want to sit beside me, right?

[Pa Joshua does not respond; he simply sits down next to Kulani.]

mboh koh mbaŋ Pa Jo but ŋgaŋ

I want to greet Pa Jo but he won't have it

Pa Joshua to Kulani

nobi ye

I'm listening to you

Kulani to Pa Joshua

notefiə ŋwo, iloŋ yiə

If you feel like listening you will, that's your trouble

Kulani to a boy nearby

ke be m-kpoko biə bə du lebin tu ke ta bebi noh

If there are more chairs, just keep bringing them

Kulani to a chicken clucking around

dzahɛ wantou noh

You too, go that way…

Abar

Buu

Source: Fir0002/Flagstaffotos

Ojong Diba, Rachel A. 2020. Nuances in language use in multilingual settings: Code-switching or code regimentation in Lower Fungom? In Pierpaolo Di Carlo & Jeff Good (eds.), *African multilingualisms: Rural linguistic and cultural diversity*, 15–28. Lanham, MD: Lexington Books.

Why is the chicken addressed in Buu?

The interchange takes place in the village of Buu

Everyone around would understand Buu

Kulani and Pa Joshua both know Buu and Abar

Why is the chicken addressed in Buu?

The interchange takes place in the village of Buu

Everyone around would understand Buu

Kulani and Pa Joshua both know Buu and Abar

Why is Pa Joshua addressed in Abar?

The interchange takes place in the village of Buu

Everyone around would understand Buu

Kulani and Pa Joshua both know Buu and Abar

Why is Pa Joshua addressed in Abar?

Pa Joshua is senior to Kulani

Kulani's father grew up in Buu

Pa Joshua is from Buu

Lower Fungom rules of language choice

1. As in other aspects of social life, the senior–junior relationship is crucial for structuring interaction
2. Language choice normally reflects the nature of the relationship holding between two people
3. The choice of a local lexicogrammatical code should be stable within an interaction
4. The code can be changed for communicative ends (e.g., for secrecy) or social ends (e.g., to signal irritation)

Missing context: Kulani was angry with Pa Joshua

Kulani's choice of Abar when speaking to Pa Joshua reflected that anger.
Kulani was more polite to the chicken than to Pa Joshua!

OLAC Record

oai:www.mpi.nl:lat_1839_00_0000_0000_0022_4B2B_8

Title:	ikaan596
	Adesonmi Obaude describes how to make pounded yam (in Ikaan)
Contributor (annotator):	Fred Adekanye
Contributor (researcher):	Funmi Chang
Contributor (speaker):	Adesonmi Obaude
Coverage:	Nigeria
	Ondo State
Format:	video/mp4
	audio/x-wav
	text/x-eaf+xml
Identifier (URI):	https://hdl.handle.net/1839/00-0000-0000-0022-4B2B-8
Is Part Of:	DoBeS archive : Ikaan
Language:	English
	Ukaan
Language (ISO639):	eng
	kcf
Publisher:	The Language Archive, Max Planck Institute for Psycholinguistics
Subject:	English language
	Ukaan language
Subject (ISO639):	eng
	kcf
Type (DCMI):	MovingImage
	Sound
	Text


```
e>
ntributor>
tributor>
ntributor>

0022-4B2B-8</dc:identifier>

/dc:publisher>

  <dc:type xsi:type="dcterms:DCMIType">Text</dc:type>
</olac:olac>
```

```
<dc:contributor xsi:type="olac:role" olac:code="speaker">Adesonmi Obaude</dc:contributor>
<dc:contributor xsi:type="olac:role" olac:code="researcher">Funmi Chang</dc:contributor>
<dc:contributor xsi:type="olac:role" olac:code="annotator">Fred Adekanye</dc:contributor>
```

How would this speaker talk to a chicken?

```
<olac:olac ...>
   <dc:title>ikaan596</dc:title>
   <dc:title>Adesonmi Obaude describes how to make pounded yam (in Ikaan)</dc:title>
   <dc:contributor xsi:type="olac:role" olac:code="speaker">Adesonmi Obaude</dc:contributor>
   <dc:contributor xsi:type="olac:role" olac:code="researcher">Funmi Chang</dc:contributor>
   <dc:contributor xsi:type="olac:role" olac:code="annotator">Fred Adekanye</dc:contributor>
   <dc:coverage>Nigeria</dc:coverage>
   <dc:coverage>Ondo State</dc:coverage>
   <dc:format>video/mp4</dc:format>
   <dc:format>audio/x-wav</dc:format>
   <dc:format>text/x-eaf+xml</dc:format>
   <dc:identifier xsi:type="dcterms:URI">https://hdl.handle.net/1839/00-0000-0000-0022-4B2B-8</dc:identifier>
   <dc:language xsi:type="olac:language" olac:code="eng">English</dc:language>
   <dc:language xsi:type="olac:language" olac:code="kcf">Ukaan</dc:language>
   <dc:publisher>The Language Archive, Max Planck Institute for Psycholinguistics</dc:publisher>
   <dcterms:ispartof>DoBeS archive : Ikaan</dcterms:ispartof>
   <dc:subject xsi:type="olac:language" olac:code="eng">English</dc:subject>
   <dc:subject xsi:type="olac:language" olac:code="kcf">Ukaan</dc:subject>
   <dc:type xsi:type="dcterms:DCMIType">MovingImage</dc:type>
   <dc:type xsi:type="dcterms:DCMIType">Sound</dc:type>
   <dc:type xsi:type="dcterms:DCMIType">Text</dc:type>
</olac:olac>
```

lameta

TatangJoyceYasho-Kung/ 0773 - lameta 2.1.2-alpha

Project | Sessions | **People**

Person
Emmanuela Chu Ngeh ...
Ernest Tang Tem Zuh
Evelyn Muh Zoo
Fidelis Tiang Njong
Filomena Meh Zai
Helen
Irene Che Bih
Isidore Nji Meh
Joseph Mbang Muh
Josephine Ngong Muh ...
Julie Tem Bei Neh
Long Che Sunfi
Martha Ndum Kah
Matthew Kah Muh
Nadesh Che Fuet
Nelson Tem Shi
Oliver

New Person

Open | Rename... | Rename for Consent | **Add Files**

Name	Type	Modified	Size
Isidore Nji Meh.person	Person	2023/12/7 15:13	1 KB

Person | Contributions | Notes | IMDI

Full Name: Isidore Nji Meh

Birth Year: 1996

Nickname:

Code:

Gender: Male

Languages

- Kung
- Mmen
- Kuk
- Aghem
- Cameroon Pidgin
- English
- French

Add Language

Description

Further information about the participant, including more precise listing of their varieties of competence, can be found in a spreadsheet accompanying this collection.

How to Contact

Based on the IMDI standard

lameta

Person | Contributions | Notes | IMDI

Full Name
Isidore Nji Meh

Birth Year
1996

Nickname

Code

Gender
Male

Languages
- Kung
- Mmen
- Kuk
- Aghem
- Cameroon Pidgin
- English
- French

Add Language

Description
Further information about the participant, including more precise listing of their varieties of competence, can be found in a spreadsheet accompanying this collection.

New Person

Based on the IMDI standard

lameta

TatangJoyceYasho-Kung/ 0773 - lameta 2.1.2-alpha

Project | **Sessions** | People

- TJ_702_0175_2016-04-14
- TJ_702_0176_2016-04-14
- TJ_702_0177_2016-04-15
- TJ_702_0178_2016-04-16
- TJ_702_0179_2016-04-17
- TJ_702_0183_2016-04-18
- TJ_702_0185_2016-04-19
- TJ_702_0187_2016-04-20
- TJ_702_0188_2016-04-28
- TJ_702_0189_2016-04-28
- TJ_702_0190_2016-04-28
- TJ_702_0191_2016-04-29
- TJ_702_0192_2016-04-30
- TJ_702_0193_2016-04-30
- TJ_702_0194_2016-04-30
- TJ_702_0195_2016-05-02
- TJ_702_0196_2016-05-02
- TJ_702_0197_2016-05-02

New Session

Open | Rename... | **Add Files**

Name	Type	Modified	Size
TJ_702_0176_2016-04-14.session	Session	2023/12/7 15:14	2 KB
TJ_702_0176_2016-04-14.MP3	Audio	2023/12/7 15:13	10 MB

Session | Contributors | Status | Notes | IMDI

ID
TJ_702_0176_2016-04-14

Date
2016-04-14
YYYY-MM-DD

Title
Elicitation of wordlist – Isidore – household items

Description
Sociolinguistic interview collected by Tatang Joyce Yasho from Isidore Nji Meh using the Sociolinguistic Interview Guide (July 2017 version) developed by the KPAAM-CAM project, a copy of which is included with this collection

Genre
Elicitation

Subgenre

Access
S: open to Subscribers

Access Explanation
Because of the sociolinguistic

Topic
interview

Keyword
individual multilingualism,

Subject Languages
Kung

Neighborhood/Town/Village
Wum, Courtyard

Working Languages
English Kung

More Fields

Field	Value
Researcher Involvement	elicited
Region	North West Region
Country	Cameroon
Continent	
Planningtype	planned
Socialcontext	

Custom Fields

Field	Value

lameta

TatangJoyceYasho-Kung/ 0773 - lameta 2.1.2-alpha

Project | Sessions | People

- TJ_702_0175_2016-04-14
- TJ_702_0176_2016-04-14
- TJ_702_0177_2016-04-15
- TJ_702_0178_2016-04-16
- TJ_702_0179_2016-04-17
- TJ_702_0183_2016-04-18
- TJ_702_0185_2016-04-19
- TJ_702_0187_2016-04-20
- TJ_702_0188_2016-04-28
- TJ_702_0189_2016-04-28
- TJ_702_0190_2016-04-28
- TJ_702_0191_2016-04-29
- TJ_702_0192_2016-04-30
- TJ_702_0193_2016-04-30
- TJ_702_0194_2016-04-30
- TJ_702_0195_2016-05-02
- TJ_702_0196_2016-05-02
- TJ_702_0197_2016-05-02

New Session

Open | Rename... | **Add Files**

Name	Type	Modified	Size
TJ_702_0176_2016-04-14.session	Session	2023/12/7 15:14	2 KB
TJ_702_0176_2016-04-14.MP3	Audio	2023/12/7 15:13	10 MB

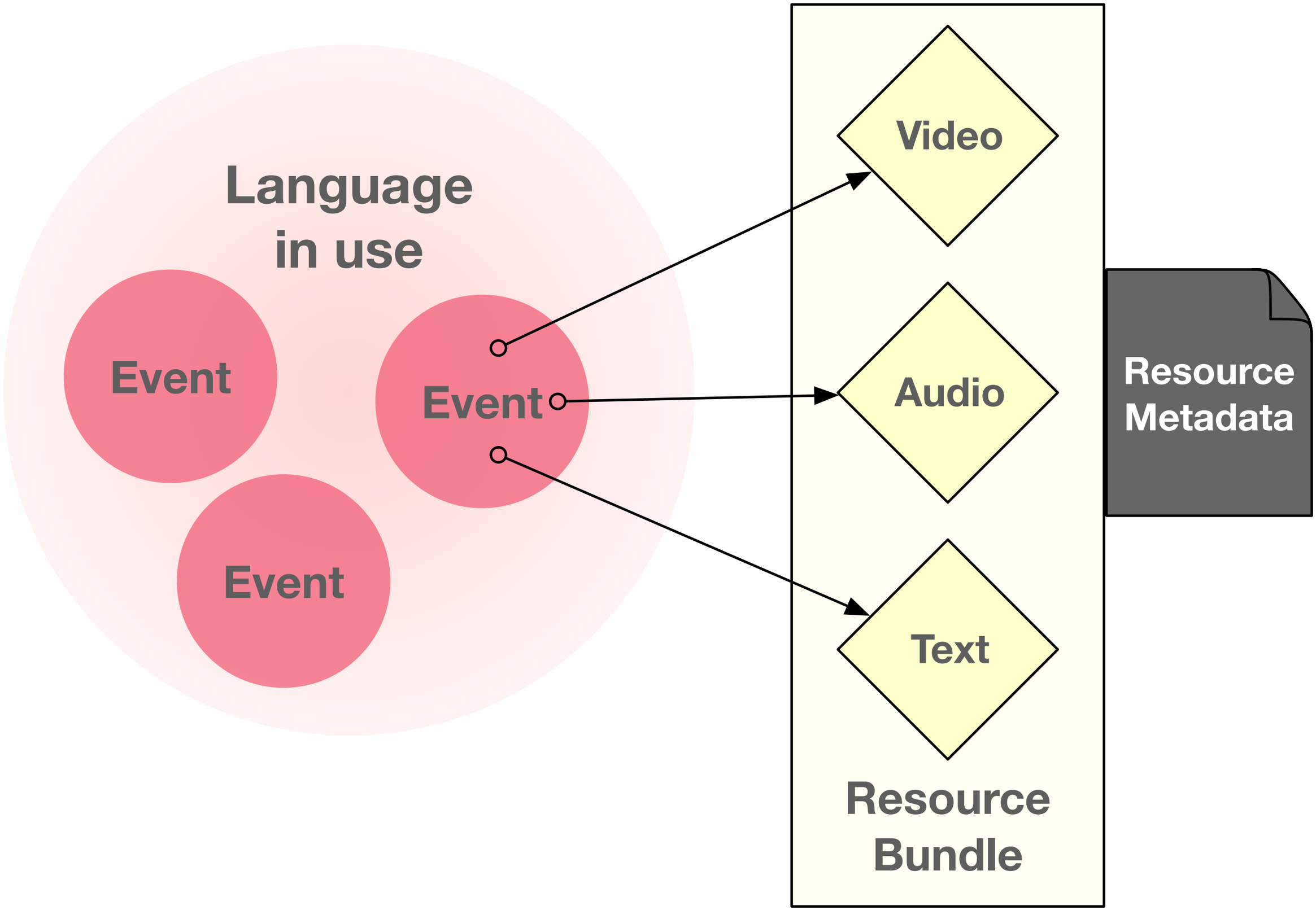
Session | **Contributors** | Status | Notes | IMDI

Name	Role	Comments
Tatang Joyce Yasho	Researcher	
Isidore Nji Meh	Consultant	

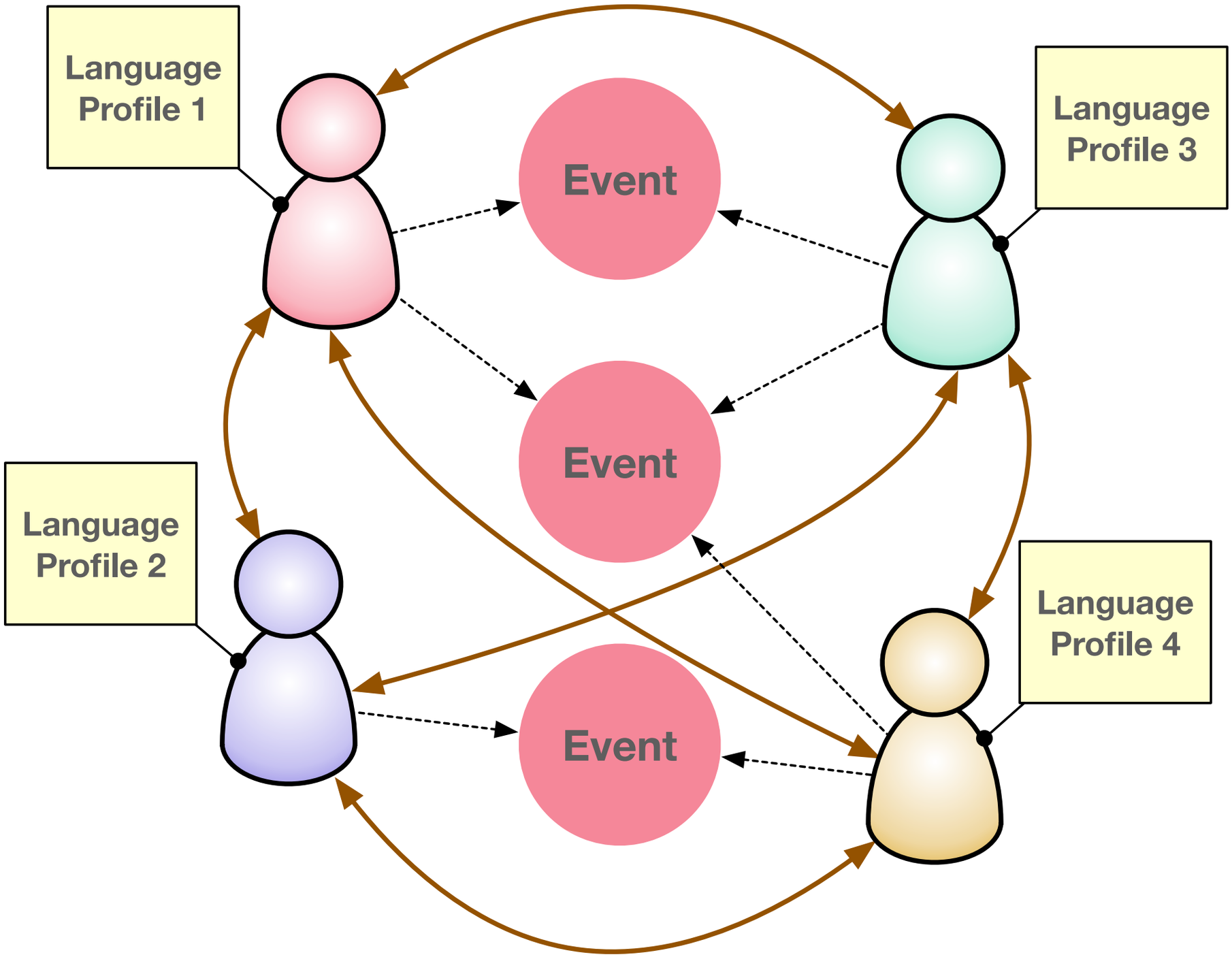
lameta

Session | **Contributors** | Status | Notes | IMDI

Name	Role	Comments	
Tatang Joyce Yasho	Researcher		
Isidore Nji Meh	Consultant		

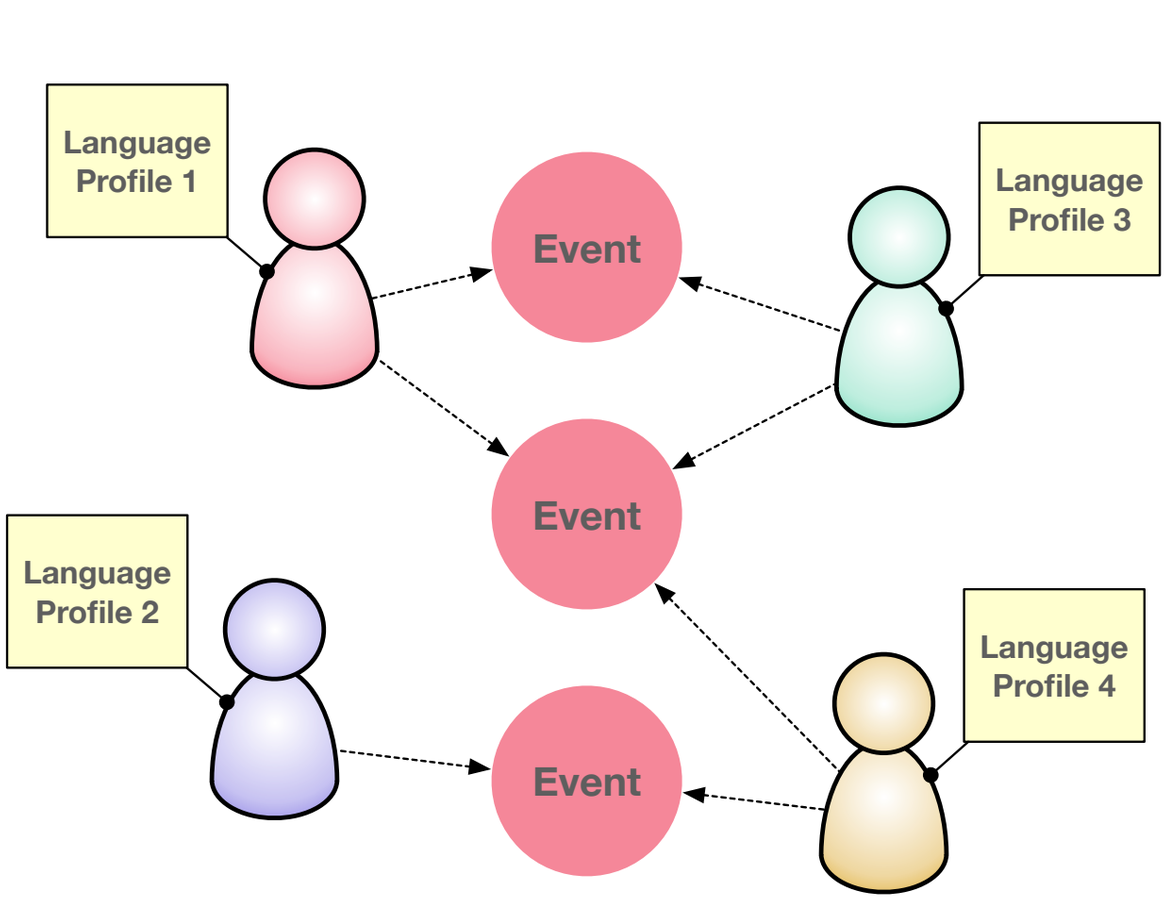


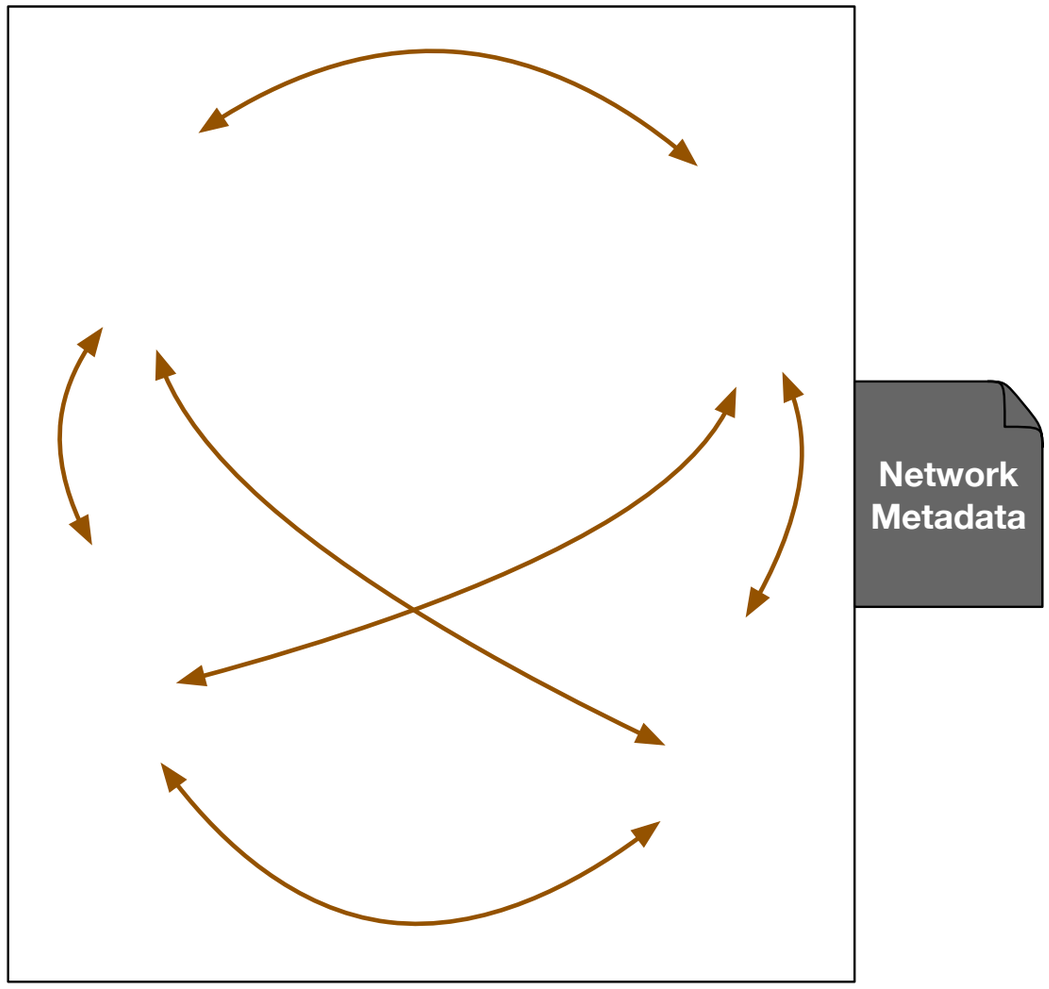
Event-centered metadata:The standard approach



Person-centered metadata: A complementary approach

Existing standards can serve as models for some extensions along these lines, e.g., FOAF RDF, TK Labels, etc.





Person-centered metadata: A complementary approach

	Audio recording →
Speaker 1	
Transcription	wàɣ lè lí hɔ́r ɓɔ́ nǐɣ ɲɛ̀rì mɔ̀
Translation	they really got upset
Language	Bisoo
Frame	gossip
Speaker 2	
Transcription	lɛ́ mâ mìndip ... ɓwám mɔ̀ ôp
Translation	give me water ... we shall start here
Language	Bisoo ... Kelleng
Frame	private ... business

Analysis of multilingual data requires annotation strategies not typical for monolingual data. Annotations won't always align this nicely!

Source: https://www.elararchive.org/dk0421/

Ngué Um, Emmanuel, Marguérite G. Makon & Célestine G. Assomo. 2020. Multilingualism as it unfolds: Language vitality in naturally occurring speech in Kelleng, a rural setting in Cameroon. In Pierpaolo Di Carlo & Jeff Good (eds.), *African multilingualisms: Rural linguistic and cultural diversity,* 53–67. Lanham, MD: Lexington Books.

	A_Transcription-txt-kuj	A_Translation-gls-en	A_Language-note	A_Participant-note-en
22	Namuge!	Il a tué!	kj	Camille
	B_Transcription-txt-kuj	B_Translation-gls-en	B_Language-note	B_Participant-note-en
23	Nifacaw!	Une bonne fois!	kj	Jean-Tomi
	D_Transcription-txt-kuj	D_Translation-gls-en	D_Language-note	D_Participant-note-en
24	Fujin beenoor ufe	Le taureau de Beenor est là!	kj	Dodo
	A_Transcription-txt-kuj	A_Translation-gls-en	A_Language-note	A_Participant-note-en
25	Ëó innuŋ?.	Où est la maman?.	gb	Camille
	B_Transcription-txt-kuj	B_Translation-gls-en	B_Language-note	B_Participant-note-en
26	Innoŋ biŋeen.	Elle est la-bas.	gb	Jean-Tomi
	C_Transcription-txt-kuj	C_Translation-gls-en	C_Language-note	C_Participant-note-en
27	Dodo!	Dodo!	-	Damace
	D_Transcription-txt-kuj	D_Translation-gls-en	D_Language-note	D_Participant-note-en
28	Père!	Père!	fr	Dodo
	C_Transcription-txt-kuj	C_Translation-gls-en	C_Language-note	C_Participant-note-en
29	Qu'est-ce qu'il y a?	Qu'est-ce qu'il y a?	fr	Damace
	D_Transcription-txt-kuj	D_Translation-gls-en	D_Language-note	D_Participant-note-en
30	Père naka mu?	Père, comment ça va?	fr w	Dodo
	C_Transcription-txt-kuj	C_Translation-gls-en	C_Language-note	C_Participant-note-en
31	Kati réunion où bien?	Peut-être une réunion où bien?	gb fr	Damace
	D_Transcription-txt-kuj	D_Translation-gls-en	D_Language-note	D_Participant-note-en
32	Waaw, réunion, liggey lep!	Oui, réunion, travail, tout!	w fr	Dodo

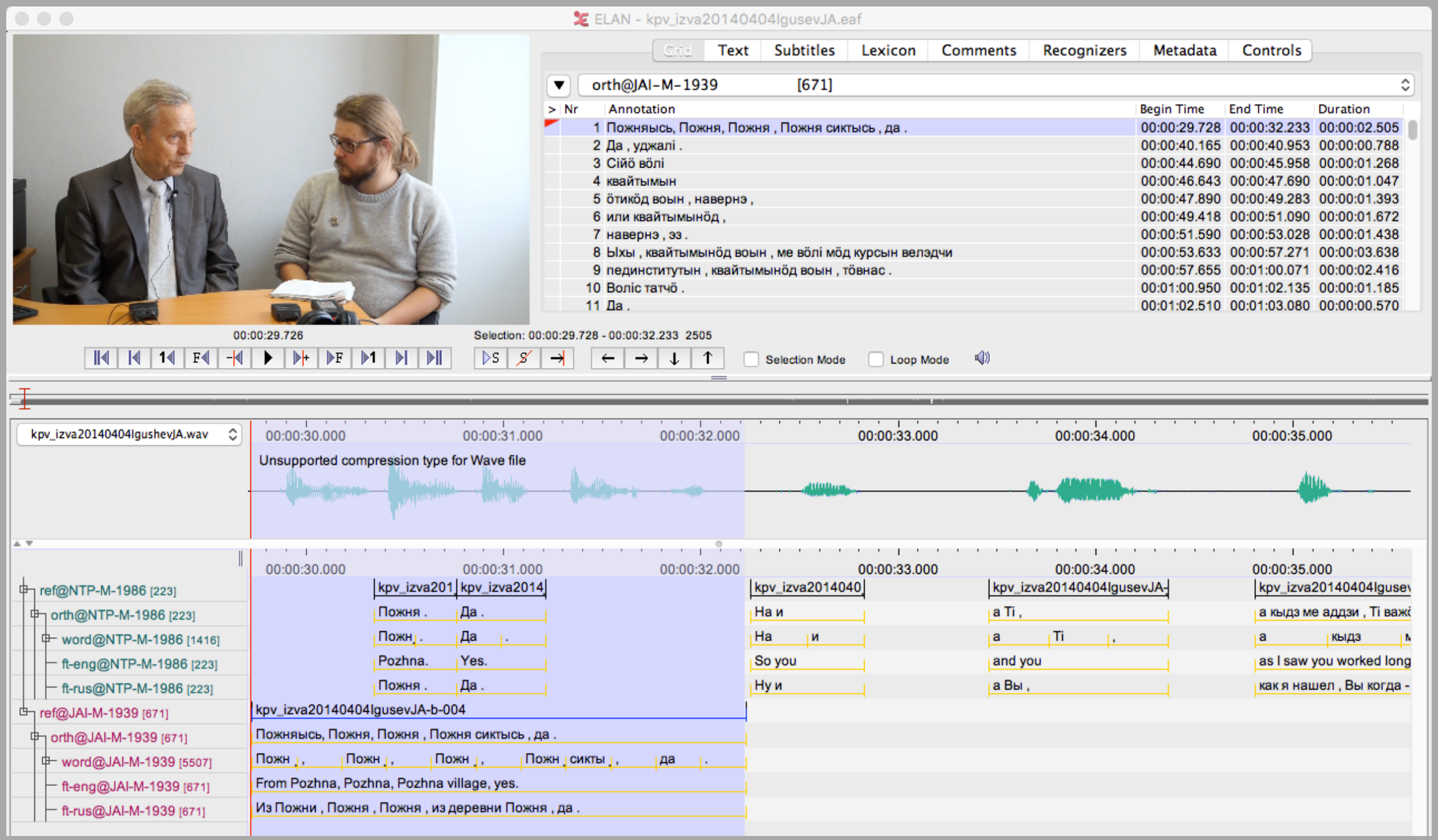
Each turn can be annotated for language choice with the possibility of combining codes

fu-jin is a mixed form with a Jola Kujirerary root and a Jola Eegima prefix

Name is not assigned to any language

Composite language tag: French–Wolof

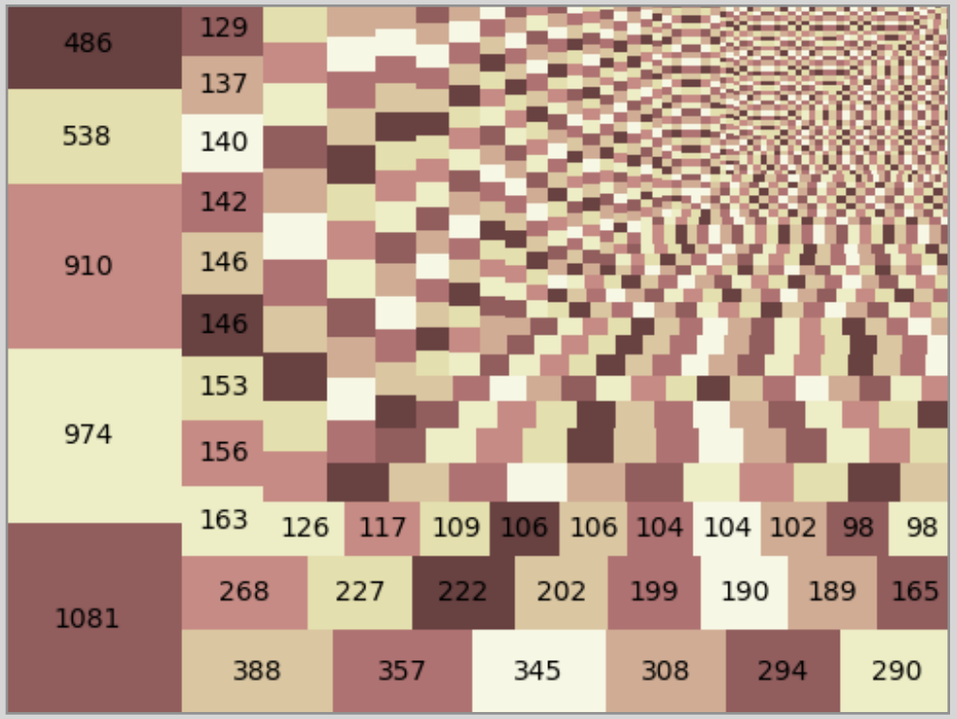
Source: Cobbinah, Alexander, Abbie Hantgan, Friederike Lüpke & Rachel Watson. 2017. Carrefour des langues, carrefour des paradigmes. In Margaret Bento Michelle Auzanneau & Malory Leclère (eds.), *Espaces, mobilités et éducation plurilingues: Éclairages d'afrique ou d'ailleurs*, 79–97. Paris: Édition des Archives Contemporaines.



Source: https://langdoc.github.io/2016-06-04-ELAN-to-R.html

ELAN is flexible enough to handle the complications associated with multilingual annotation.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<ANNOTATION_DOCUMENT>
    <TIME_ORDER>
        <TIME_SLOT TIME_SLOT_ID="ts1" TIME_VALUE="740"/>
        <TIME_SLOT TIME_SLOT_ID="ts2" TIME_VALUE="1860"/>
        <TIME_SLOT TIME_SLOT_ID="ts3" TIME_VALUE="3718"/>
        ...
    </TIME_ORDER>
    <TIER TIER_ID="ref@DAM" PARTICIPANT="Dambar Baram" ANNOTATOR="KP" LINGUISTIC_TYPE_REF="ref">
            <ANNOTATION>
                <ALIGNABLE_ANNOTATION ANNOTATION_ID="ann0" TIME_SLOT_REF1="ts1" TIME_SLOT_REF2="ts2">
                    <ANNOTATION_VALUE>. 001</ANNOTATION_VALUE>
                </ALIGNABLE_ANNOTATION>
            </ANNOTATION>
            <ANNOTATION>
                <ALIGNABLE_ANNOTATION ANNOTATION_ID="ann8" TIME_SLOT_REF1="ts3" TIME_SLOT_REF2="ts4">
                    <ANNOTATION_VALUE>. 002</ANNOTATION_VALUE>
                </ALIGNABLE_ANNOTATION>
            </ANNOTATION>
            ...
    </TIER>
    <TIER TIER_ID="ut@DAM" PARTICIPANT="Dambar Baram" ANNOTATOR="KP" LINGUISTIC_TYPE_REF="ut" PARENT_REF="ref@DAM">
            <ANNOTATION>
                <REF_ANNOTATION ANNOTATION_ID="ann1" ANNOTATION_REF="ann0">
                    <ANNOTATION_VALUE>əbə</ANNOTATION_VALUE>
                </REF_ANNOTATION>
            </ANNOTATION>
            <ANNOTATION>
                <REF_ANNOTATION ANNOTATION_ID="ann9" ANNOTATION_REF="ann8">
                    <ANNOTATION_VALUE>kunəi pudza tukle hon lə məlak</ANNOTATION_VALUE>
                </REF_ANNOTATION>
            </ANNOTATION>
            <ANNOTATION>
                <REF_ANNOTATION ANNOTATION_ID="ann36" ANNOTATION_REF="ann35">
                    <ANNOTATION_VALUE>hidi hudi pudza tukle alam alam wa lakle əbə</ANNOTATION_VALUE>
                </REF_ANNOTATION>
            </ANNOTATION>
            ...
    </TIER>
    <LINGUISTIC_TYPE LINGUISTIC_TYPE_ID="ref"/>
    <LINGUISTIC_TYPE LINGUISTIC_TYPE_ID="ut" CONSTRAINTS="Symbolic_Association"/>
    <LINGUISTIC_TYPE LINGUISTIC_TYPE_ID="txd" CONSTRAINTS="Symbolic_Association"/>
    <LINGUISTIC_TYPE LINGUISTIC_TYPE_ID="tx" CONSTRAINTS="Symbolic_Subdivision"/>
    <LINGUISTIC_TYPE LINGUISTIC_TYPE_ID="mb" CONSTRAINTS="Symbolic_Subdivision"/>
    <LINGUISTIC_TYPE LINGUISTIC_TYPE_ID="ge" CONSTRAINTS="Symbolic_Association"/>
    <LINGUISTIC_TYPE LINGUISTIC_TYPE_ID="ft" CONSTRAINTS="Symbolic_Association"/>
</ANNOTATION_DOCUMENT>
```



But, ELAN is not editable by hand, and, without standards, the costs for data interoperability will be quite high.

Kilu von Prince and Sebastian Nordhoff. 2020. An empirical evaluation of annotation practices in corpora from language documentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2778–2787. Marseille: European Language Resources Association.

CLDF and multilingual (text) data: Observations

❶ Multilingual text analysis requires access to detailed linguistic metadata

❷ Multilingual text data creates challenges for language identification

❸ Multilingual text annotations present complications that will challenge the CLDF principles, but existing alternatives are already problematic