

NINJAL, 2026.3.14

ELAN/eaftox/InTex/InDIGO Workshop

Michinori Shimoji

smz@kyudai.jp

ブラウザでご覧になりたい方はこちらから

このワークショップでは

- データを整える：生データから最低限の注釈データへ
- データを守る：注釈済みデータを**適切に**保存するまで
- データを生かす：適切に保存したその先にできること

今日使うデータ一式はここからダウンロードできます。

	00:00:02.000	00:00:03.000	00:00:04.000	00:00:05.000					
text [4]	nkjaan=du anna=tu uja=tu tamamiga=ti asi ffa-gama=nu u-tar=ca								
morph [49]	nkjaan	du	anna	tu	uja	tu	tamamig	ti	asi
gloss [49]	昔	FOC	母親	COM	父親	COM	タマミガ	QUOT	言う
trans [4]	昔々、母親と父親とタマミガという子供がいたんだとさ。								

このワークショップでは

- データを整える：録音に最低限の注釈をつけるまで
- データを守る：注釈済み談話データを**適切に**保存するまで
- データを生かす：適切に保存したその先にできること（紹介）

今日使うデータ一式はここからダウンロードできます。

	00:00:02.000	00:00:03.000	00:00:04.000	00:00:05.000					
text [4]	nkjaan=du anna=tu uja=tu tamamiga=ti asi ffa-gama=nu u-tar=ca								
morph [49]	nkjaan	du	anna	tu	uja	tu	tamamig	ti	asi
gloss [49]	昔	FOC	母親	COM	父親	COM	タマミガ	QUOT	言う
trans [4]	昔々、母親と父親とタマミガという子供がいたんだとさ。								

録音について

- マイクは個人的に**AKG c520**がおすすめ。
- 録音機材はXLR対応で。
2025年段階では**Zoom H6/Marantz PMD661MKIII**などが手頃



Workshop 1:

データを整える

録音に最低限の注釈をつけるまで

ELANで注釈をつけてみよう

- exercise.wav (日本語による1分の録音済み談話, frog story)
- temp3.etf
- temp4.etf
- tempwb.etf

Ex 1 冒頭10文節に対して形態素に分け、グロスをつける練習

Ex 2 形態素融合に対応するには

Ex 3 仮名を使ったとき、注釈をどうするの？

Ex 4 品詞層 (POS層) も含める方針：日琉スタンダードにしたい

Workshop 2: データを守る

注釈済み談話データを**適切に**保存するまで

データをどこに保存してありますか？

マイドライブ > 2025_椎葉調査 > 20251027_day1 ▾ 🗄

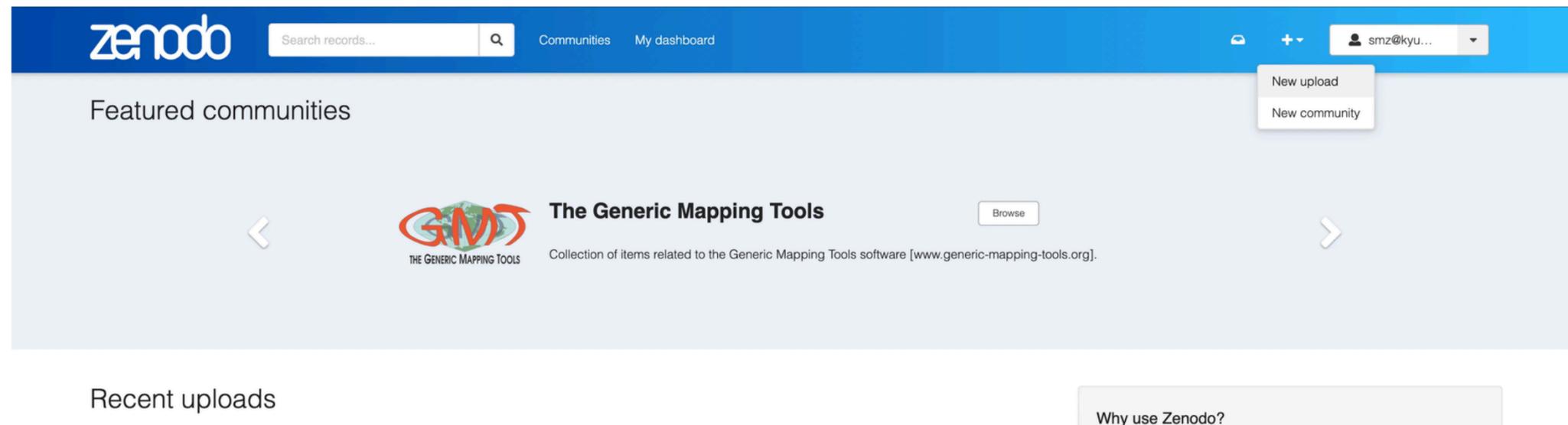
種類 ▾ ユーザー ▾ 最終更新 ▾ ソース ▾

📄 ストレージの 70% を使用しています 空き容量がなくなると、ファイルの作成、編集、アップロードができなくなります。200 GB の保存容量を月額 ¥440 でご利用いただけます。

名前	オーナー	更新日時 ↓
 20251027_shiiba_omae_ .eaf 🧑	 S	11月21日
 20251113.mp4 🧑	 M 自分	11月13日 自分
 vocabtemp.etf 🧑	 M 自分	11月13日 自分
 20251027_shiiba_omae_ .eaf 🧑	 M 自分	11月13日 自分
 ninjalvocab.xlsx 🧑	 M 自分	11月13日 自分
 20251027_shiiba_omae_metadata.xlsx 🧑	 M 自分	10月30日 自分
 20251027_shiiba_omae_ WAV 🧑	 M 自分	10月27日 自分

言語ドキュメンテーションの鉄則

- ローカルとクラウドで二重保存
 - データの散逸を防ぐ
 - データの更新状況の透明化
- オープンアクセスのリポジトリを活用する



3つのステップ

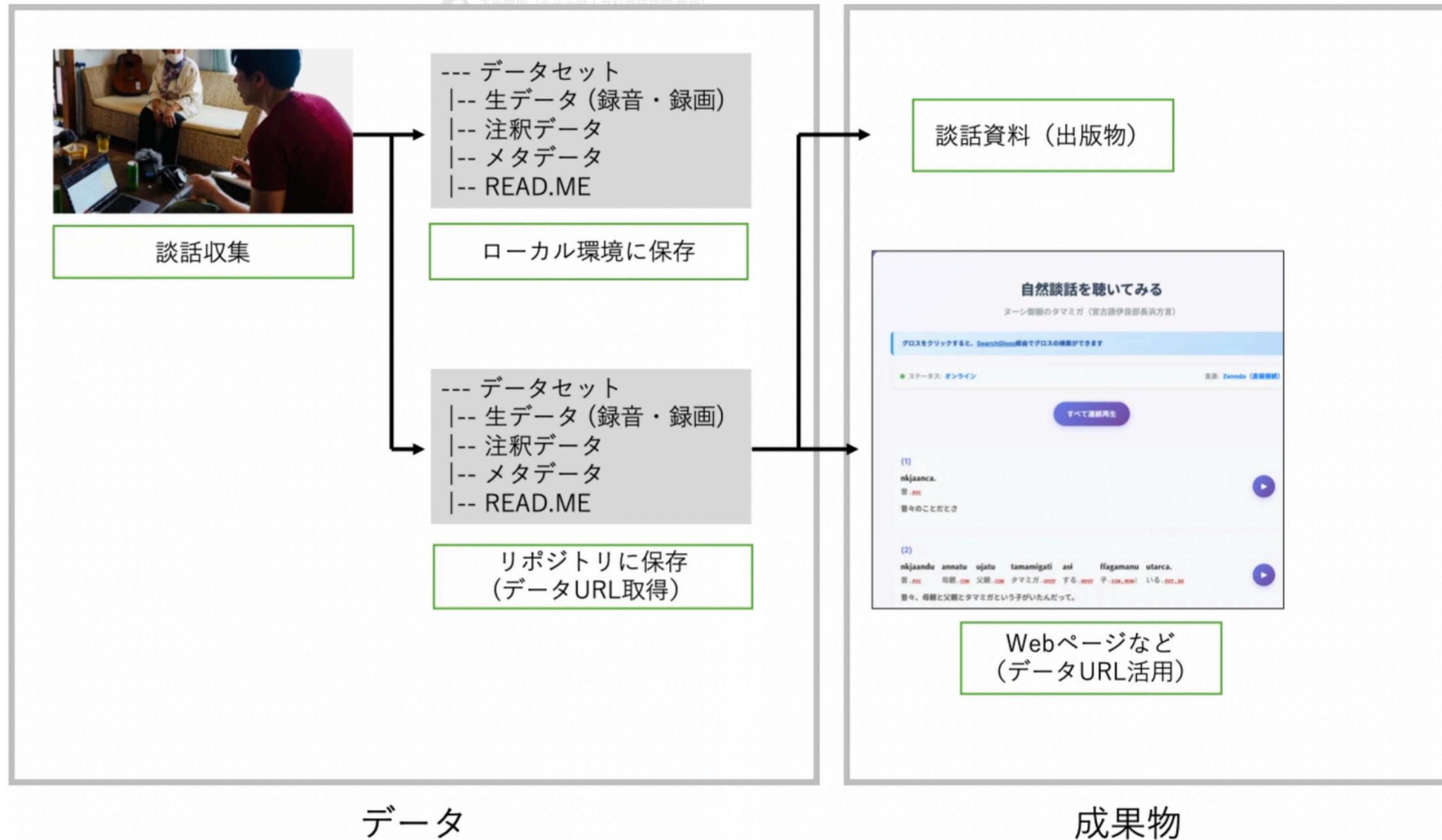
note記事「消滅危機言語の記録保存（特に談話データを例に）」
詳しくは[ここで](#)読めます

Step 1 絶対にすべきこと：データの永続性。自分のローカル環境（パソコンのHDとかSDカード保存とか）に加えて、**ウェブ上の安定したリポジトリ**にも関連データを全て置く。関連データとは、生データ（録音、録画）、書き起こしデータ（注釈付き）、メタデータ（話者情報、注釈者情報など）

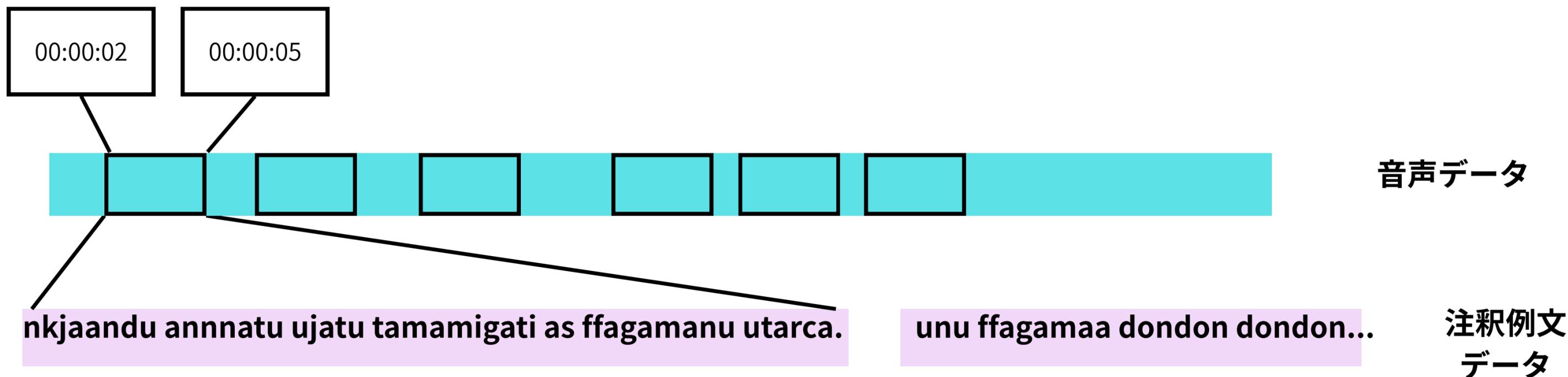
Step 2 望ましいこと：少なくともものちの研究者/コミュニティが利用しやすいような仕組みを整えておく。後の修正が入る場合のバージョン管理の徹底。個々のデータに対するデジタル資料の固有識別番号であるDOI（Digital Object Identifier）の取得など。

Step 3 できると素晴らしいこと：コミュニティが、自らの言語復興や継承言語教育のリソースに使いやすくする。目的特化型のWebページを作り、誰でもわかりやすい方法で談話資料を公開する。談話の例文ごとに音声も聞けるような仕組みなどがあると素晴らしい。

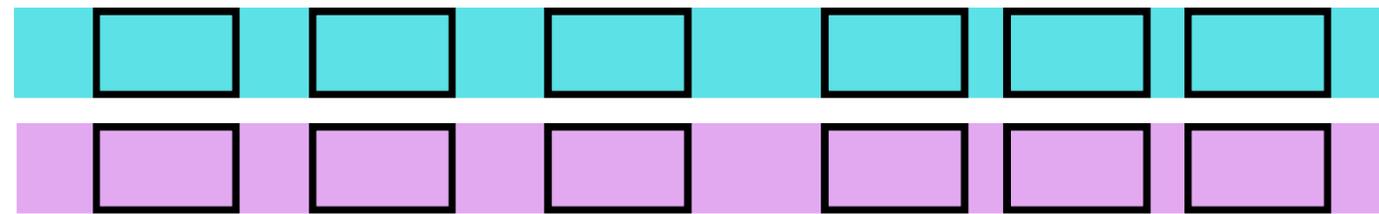
3つのステップ：具体的な実現手順



ELANで生み出される2つのデータ



	2.000	00:00:03.000	00:00:04.000	0			
text [4]	nkjaan=du annnatu ujatu tamamigati as fflagamanu utarca.						
morph [49]	nkjaan	du	anna	tu	uja	tu	tam
gloss [49]	昔	FOC	母親	COM	父親	COM	タマ
trans [4]	昔々、母親と父親とタマミガという子供がいたんだとさ。						



eaftox

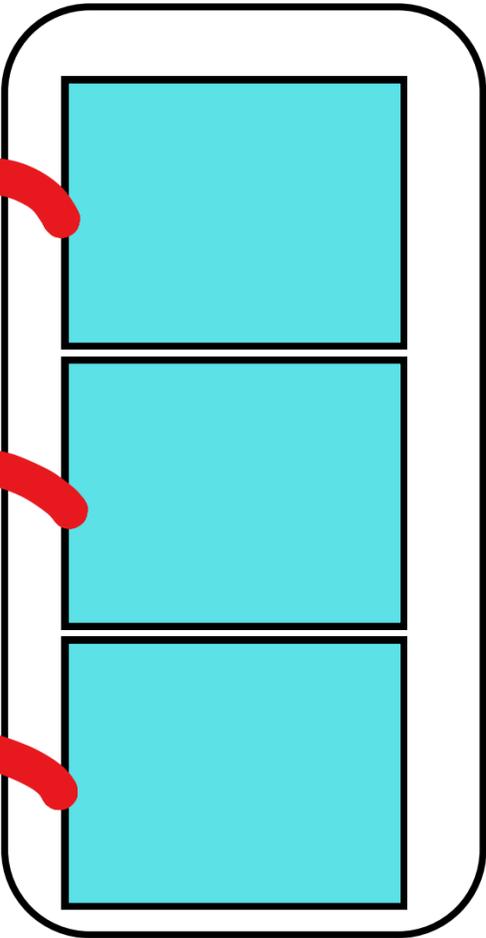
注釈例文ファイル
txt

音声ファイル
wav

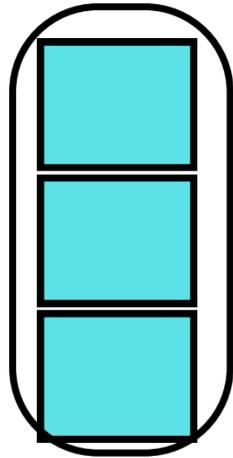
nkjaan=du annna=tu uja=tu tamamiga=ti as ffa-gama=nu u-tar=ca
昔=FOC 母親=COM 父親=COM タマミガ=QUOT 子-DIM=NOM いる-PST=HS
「昔々、母親と父親とタマミガという子がいたんだとさ」

unu ffa-gama=a dondon dondon aparagi midum-masii sudac-i-i
その 子-DIM=TOP ONM ONM 美しい 女-CIRCM 育つ-THM-SEQ
「その子はどんだんどんだん美人に育っていき」

mura=kara=mai icban aparagi-munu=n nar-i-i nar-i-u-tar=ca
村=ABL=ADD 一番 美しい-PRED=DAT なる-THM-SEQ なる-THM-PROG-PST=HS
「村でも一番の美人になっていったんだとさ」



注釈内容と音声を関連づけて別々に保存



nkjaan=du annna=tu uja=tu tamamiga=ti as ffa-gama=nu u-tar=ca
昔=FOC 母親=COM 父親=COM タマミガ=QUOT 子-DIM=NOM いる-PST=HS
「昔々、母親と父親とタマミガという子がいたんだとさ」

unu ffa-gama=a dondon dondon aparagi midum-masii sudac-i-i
その子-DIM=TOP ONM ONM 美しい女-CIRCM 育つ-THM-SEQ
「その子はどんだんどん美人に育っていき」

mura=kara=mai icban aparagi-munu=n nar-i-i nar-i-u-tar=ca
村=ABL=ADD 一番 美しい-PRED=DAT なる-THM-SEQ なる-THM-PROG-PST=HS
「村でも一番の美人になっていったんだとさ」

Name	Size	Download all
001_nkjaanca.wav md5:a447944b6e593d49219e5af8ff6a3f8a ⓘ	134.4 kB	Preview Download
002_nkjaandu_annatu_ujatu_tamam.wav md5:5a31adbea4bf5d91c1ee8363ebec87fe ⓘ	578.0 kB	Preview Download
003_unu_ffa-gamaa_dondon_dondon_a.wav md5:f6fa8cc15882beba7dd3d5d6e7458977 ⓘ	505.0 kB	Preview Download
004_murakaramai_icban_aparagi-mu.wav md5:722a831a93e12a1a5d19e4172be42f90 ⓘ	651.9 kB	Preview Download
sample.wav md5:325005ab85c3099f551cd2898347c272 ⓘ	2.0 MB	Preview Download
sample3.txt md5:505c6c8b080c11b330ba825cf2e05012 ⓘ	878 Bytes	Preview Download

個別保存するメリット絶大（後述）

The diagram illustrates the process of saving individual files. On the left, a folder icon is connected by a blue arrow to a file list table. Another blue arrow points from a text snippet to the file list. The file list contains the following items:

Name	Size	Download all
001_nkjaanca.wav md5:4479446e693949219e5a8f9630a	134.4 kB	Preview Download
002_nkjaandu_anna_ujatu_tamam.wav md5:5a37a8ba405d91e7e6303e3e0c37e	578.0 kB	Preview Download
003_unu_lla-gamaa_dondon_dondon_a.wav md5:9f8d0c1582b2a7a7a5d5d5e745977	505.0 kB	Preview Download
004_murakaramai_iban_aparagi-mu.wav md5:722a831a03a12a1a519a4172e4290	651.9 kB	Preview Download
sample.wav md5:35005a85c309951c02898347272	2.0 MB	Preview Download
sample3.txt md5:505c6c85080c119330ca825c70a0512	878 Bytes	Preview Download

The text snippet shows a line of text with a purple highlight: `nu-lla-gamaa-dondon-dondon-anna-ujatu-tamam`. Below it, a Japanese translation is visible: 「そのはだんどんどん美人に言ってい」.

Zenodo

ヌーシ御嶽のタマミガ

宮古語伊良部長浜方言

グロスをクリックすると、和訳が表示されます (SearchGlossに基づくグロス辞書を内蔵)

● ステータス: 準備完了 例文数: 54

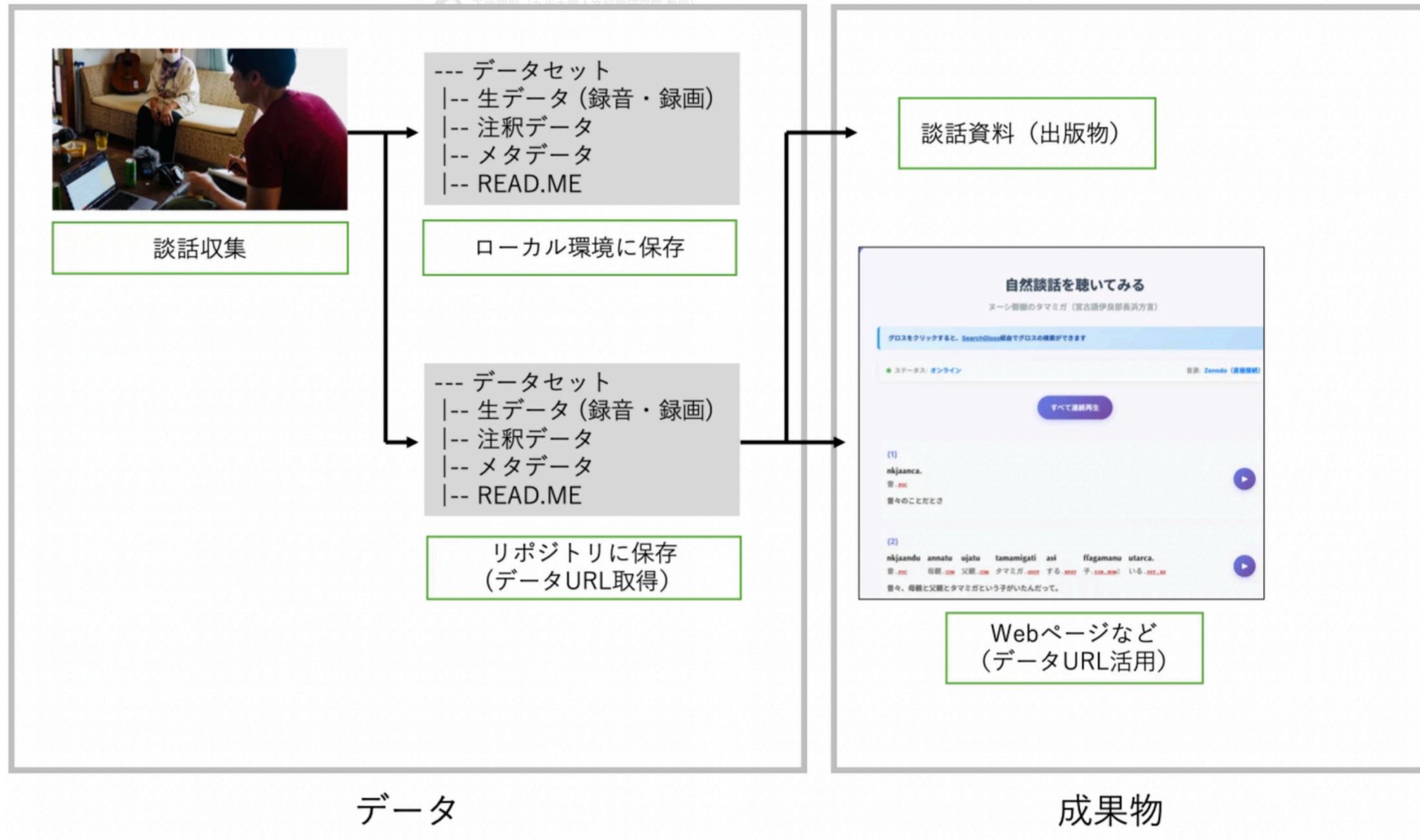
▶ 全体音声を再生 ↓ HTMLをダウンロード

(1)
nkjaan=ca.
昔=HS
昔々のことだとさ

(2)
nkjaan=du anna=tu uja=
昔=FOC 母親=COM 父親=
昔々、母親と父親とタマミガと

文ごとに再生可能な
Webページ

Zenodoにデータをアップしてみよう

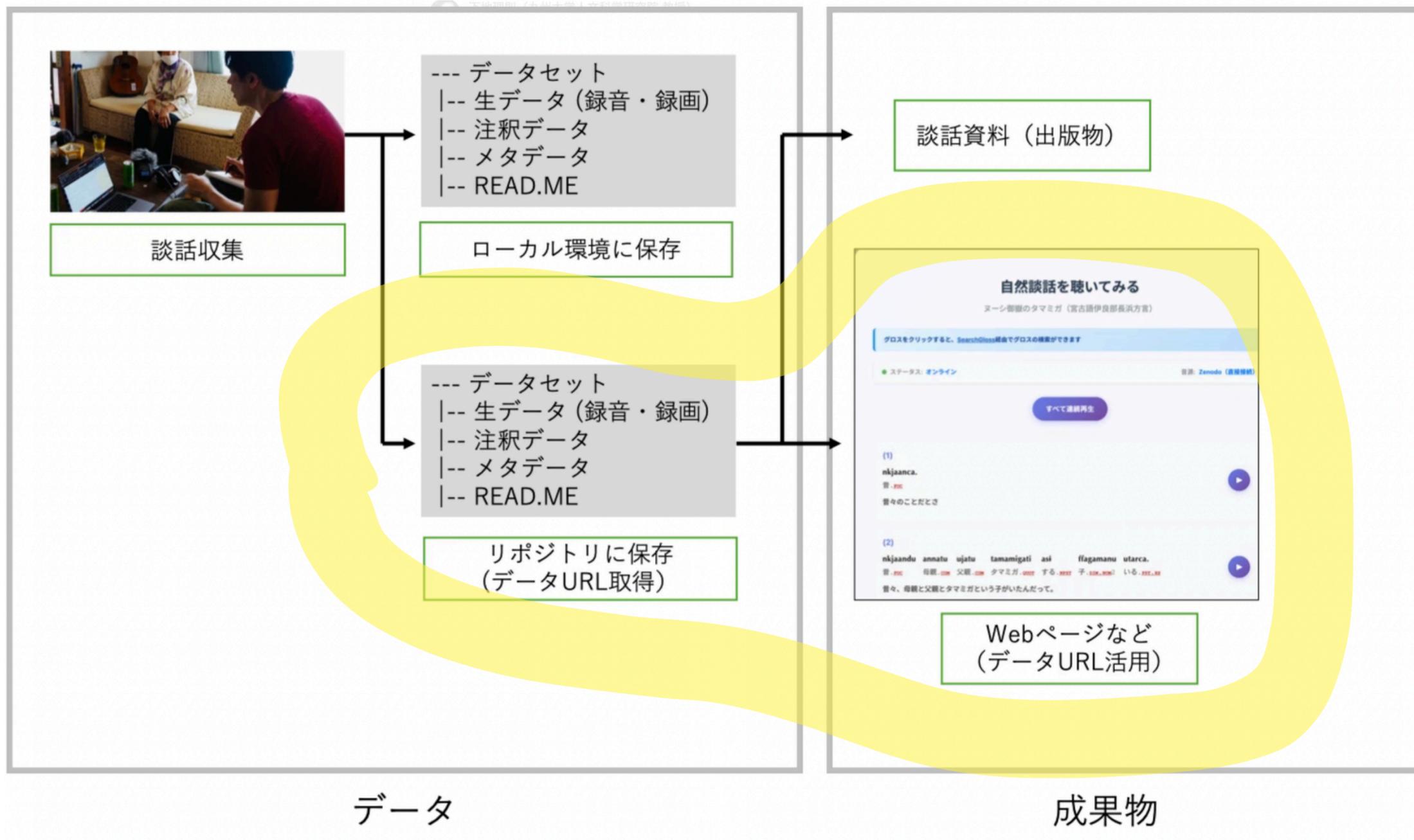


Workshop 3:

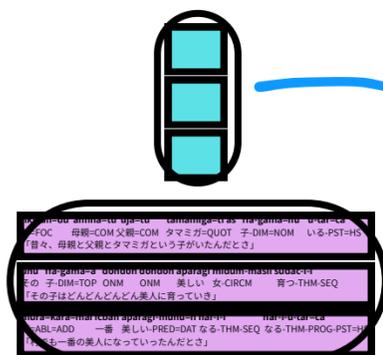
データを生かす

適切に保存したその先にできること

Workshop 3でやること



InTEEx: zenodo経由でWebページ作成



Name	Size	Download all
001_nkjaanca.wav md5:4479446e693949219e5a8f9630a	134.4 kB	Preview Download
002_nkjaandu_anna_tu_uja_tu_tamam.wav md5:5a37a8ba485d91e7e0303e30e037e	578.0 kB	Preview Download
003_unu_ija_gamaa_dondon_dondon_a.wav md5:9f80c1682b2a7a5d5d5e7459977	505.0 kB	Preview Download
004_murakaramai_iban_aparagi-mu.wav md5:72a831a05a12a1a519a4172e4290	651.9 kB	Preview Download
sample.wav md5:35005a85c309951c02898347272	2.0 MB	Preview Download
sample3.txt md5:505c6c85080c119330ca825c705012	878 Bytes	Preview Download

eaftox

Zenodo

ヌーシ御嶽のタマミガ

宮古語伊良部長浜方言

グロスをクリックすると、和訳が表示されます (SearchGlossに基づくグロス辞書を内蔵)

● ステータス: 準備完了

例文数: 54

▶ 全体音声を再生

↓ HTMLをダウンロード

(1)
nkjaan=ca.
昔=HS
昔々のことだとさ

(2)
nkjaan=du anna=tu uja=tu tar... u-tar=ca.
昔=FOC 母親=COM 父親=COM タ... いる-PST=HS
昔々、母親と父親とタマミガという子が

InTEEx

pos層を整備しておく 可能性が広がる

	.500	00:00:16.000	00:00:16.500	00:00:17.000	00:00:17.500	00:00:18.000	00:00:18.500	0	
	nkjaan=du anna=tu uja=tu tamamiga=ti asi ffa-gama=nu u-tar=ca.								
text	nkjaan	du	anna	tu	uja	tu	tamamiga	ti	asi
gloss	昔	FOC	母親	COM	父親	COM	タマミガ	QUOT	いう
pos	N	ISP	N	CP	N	CP	PN	CJP	V
trans	昔々、母親と父親とタマミガという子がいたんだって。								

pos層を整備しておく と可能性が広がる

- ELANのデータから簡易電子辞書を作る際に便利 →link
- ELANのデータからPOS一覧を作り、簡易文法スケッチを作成する足掛かりになる
(ELAN → json → Pythonスクリプトで生成) →link
- 機械学習でグロス自動化を実装する上で、精度の高い教師データになる (下地2026) →link

おわりに

このワークショップの振り返り

WS1 データを整える：録音に最低限の注釈をつけるまで

WS2 データを守る：注釈済み談話データを**適切に**保存するまで

WS3 データを生かす：適切に保存したその先にできること

- WS2, 3の解説動画は**ここから**ダウンロードできます（九大の下地ゼミ院演習録画）
- eaftoxの解説記事と動画は**こちらから**（note記事）
- InTEExの解説記事と動画は**こちらから**（note記事）
- 消滅危機言語のドキュメンテーションの概説記事は**こちらから**（note記事）
- グロス付けのツール全般は**こちらから**（下地理則研究室HP内）