

Introduction to Open Text Collections

Sebastian Nordhoff, Mandana Seyfeddinipur & Christian Döhler

Berlin-Brandenburgische Akademie der Wissenschaften



open
text
collections



- ↪ project started September 2023
- ↪ funded through DFG for 3 years
- ↪ after the funding period the project will be supported via Language Science Press
- ↪ 5 regional boards:
 - ↪ Africa, Eurasia, Caucasus, Papunesia, South America,
 - ↪ (Meso-America?)
- ↪ should produce about 8 collections/year with a total of 40k IMT tokens/year.



History of the project

- first idea 2017
- first grant proposal submitted May 2018
- rejected May 2019
 - no connection to existing archiving institutions
 - too few funds requested
- new proposal 2022
 - attach to BBAW
 - ask for one extra postdoc position
- granted February 2023
- start September 2023



- The **Berlin-Brandenburg Academy of Sciences and Humanities** (BBAW) was founded in 1700
- German and foreign dictionaries
- critical editions of texts from antiquity to modern times
- digital presentation of canonical works from various fields of the humanities.
- Leading player in the field of Digital Humanities
- good overlap in domain and technology with Open Text Collections
- e.g. Corpus of egyptological texts, which allows lexical queries across all epochs of Egyptian



- The **Endangered Language Documentation Project (ELDP)** gives grants for people to document endangered languages
 - ELDP supports the documentation and preservation of endangered languages through granting, training and outreach activities.
 - The collections compiled through our funding are freely accessible at the Endangered Languages Archive.
- **The Endangered Language Archive (ELAR)** hosts archival data for many endangered languages
 - digital collections, including audio and video recordings, of endangered languages
 - heterogeneous
 - eclectic
 - no clear querying path
- → tell depositors to collect their data in a way that can easily be fed into pipelines



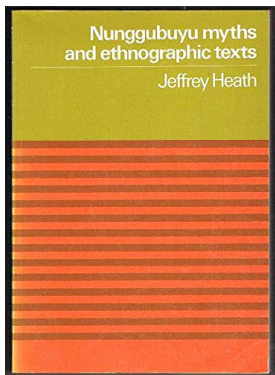
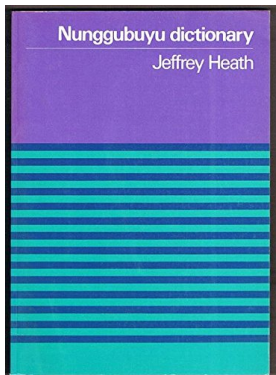
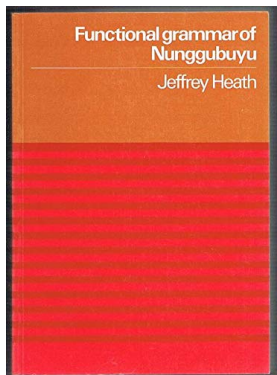
What is a text collection?

- ↪ **curated**: a selection of texts from a language, e.g. from a documentation project
 - ↪ during the funding period, we focus on monologues
- ↪ **contextualized**: introduction to the speech community, language, recording methods, speakers, etc.
- ↪ **edited**: contributors may choose to edit out false starts, pauses, self-corrections, etc.
 - ↪ the texts will not be naturalistic
 - ↪ not time-aligned with audio-visual footage
- ↪ **transparent**: good provenance
 - ↪ texts will be linked to the original recordings (or scans) in an archive
 - ↪ editing decisions will be documented
- ↪ **accessible**: open-access, interoperable
- ↪ **glossed**: interlinearized text following the Leipzig Glossing Rules



Existing text collections

- Most grammars have glossed texts in the appendix
- Goes back to Franz Boas' idea of grammar, dictionary and text as a trilogy of language documentation
- Jeffrey Heath's trilogy of Nunggubuyu:





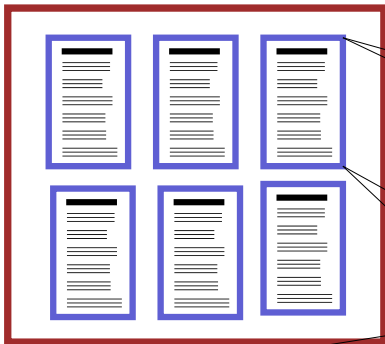
Existing text collections

- NATAMX contains 12 volumes of legacy text collections published between 1976-1980.
 - Access has to be purchased
 - The texts are in pdf format only
- TIPAA has texts in the languages of the Americas
 - Not open access
 - Not available in a structured format
- TILP has texts in a variety of languages from the Pacific.
 - The texts are free to download, but have no licence information.
 - The texts are in pdf format only.
- PANGLOSS features some texts, but also a lot of other materials.
 - When accessing materials about a certain language, you cannot be sure what you will find.
- OTC will be **open access** and in a **structured format**



What is a text collection?

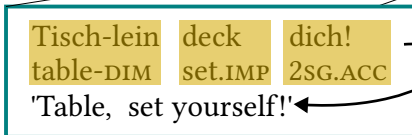
Collection

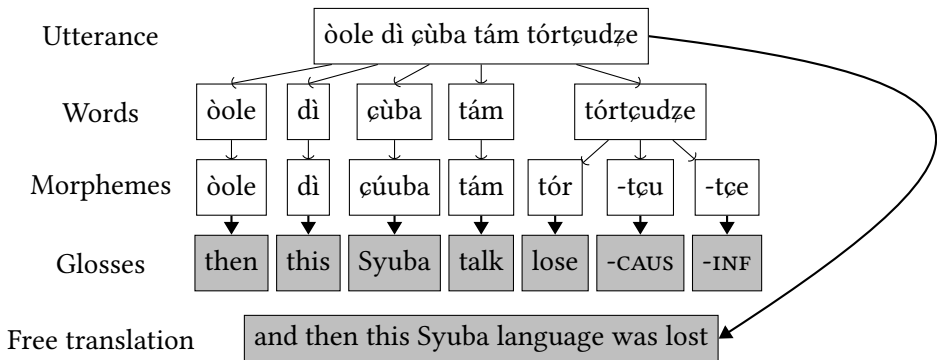


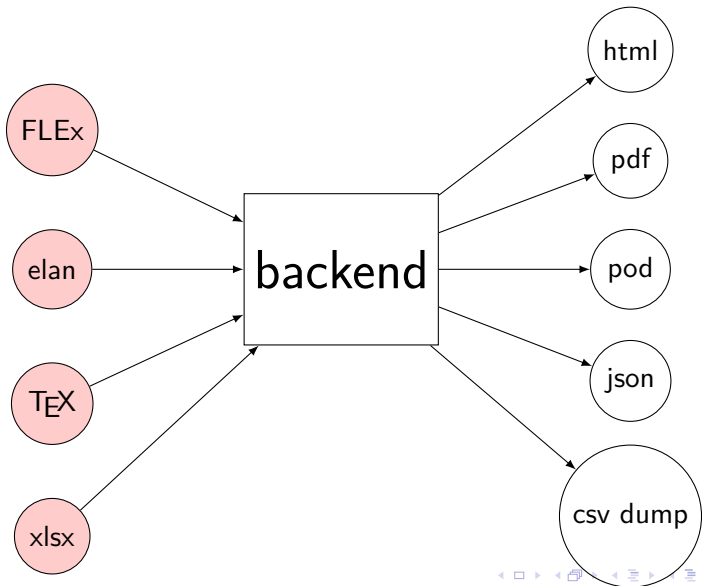
Text



Sentence

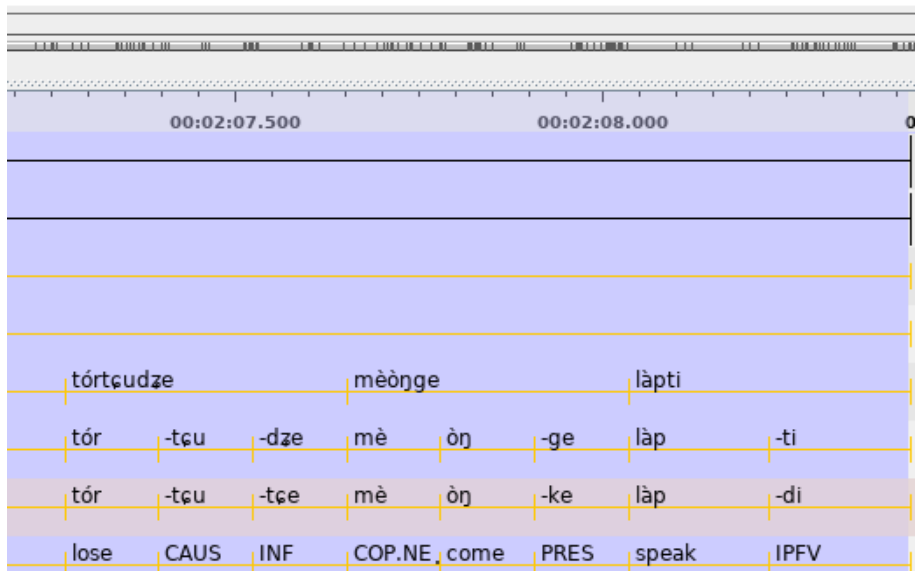









Input formats: ELAN





Input formats: ELAN

```
<ANNOTATION>
<REF_ANNOTATION_ANNOTATION_ID="ann1564_flexid_36991673-5868-4983-ad94-377a7e72d0f7" ANNOTATION_REF="ann1542_flexid_6407c08d-440d-4e72-9a24-93dafd708b79"
PREVIOUS_ANNOTATION="ann1558_flexid_6d48270b-d00c-41e2-b4b6-86fb026de5d5">
  <ANNOTATION_VALUE>tórtcudze</ANNOTATION_VALUE>
</REF_ANNOTATION>
</ANNOTATION>
<ANNOTATION>
<REF_ANNOTATION_ANNOTATION_ID="ann1579_flexid_41ffa0f4-476f-4c4d-b95d-ffb6f0688ad4" ANNOTATION_REF="ann1542_flexid_6407c08d-440d-4e72-9a24-93dafd708b79"
PREVIOUS_ANNOTATION="ann1564_flexid_36991673-5868-4983-ad94-377a7e72d0f7">
  <ANNOTATION_VALUE>méõjge</ANNOTATION_VALUE>
</REF_ANNOTATION>
</ANNOTATION>
<ANNOTATION>
<REF_ANNOTATION_ANNOTATION_ID="ann1593_flexid_34f28fad-f29c-4a90-8afa-846857914b39" ANNOTATION_REF="ann1542_flexid_6407c08d-440d-4e72-9a24-93dafd708b79"
PREVIOUS_ANNOTATION="ann1579_flexid_41ffa0f4-476f-4c4d-b95d-ffb6f0688ad4">
  <ANNOTATION_VALUE>lápti</ANNOTATION_VALUE>
</REF_ANNOTATION>
</ANNOTATION>
```

 ingestion via e1dpy python library



Input formats: FLEEx

Kalaba - FieldWorks Language Explorer

File Edit View Data Insert Format Tools Parser Window Help

Texts & Words

- Interlinear Texts
- Concordance
- Word List Concordance
- Word Analyses
- Bulk Edit Wordforms
- Statistics

Texts

Title

Show All

My Green Mat

Text

Title

Info Baseline Gloss Analyze Tagging Print View Text Chart


1.1	Word	pus	yalola	nihimbilira	.
	Word Gloss	green	I see		
	Word Cat.	mod	V		
1.2	Word	nihimbilira	pus	yalola	.
	Word Gloss	I perceive	green	my mat	
	Word Cat.	V	mod	N	
1.3	Word	hesyla	nihimbilira	.	
	Word Gloss	***	I see		
	Word Cat.	***	V		

25/Mar/2011 Queue: (-/-) No Parser Loaded Sorted by Title 1/1



Input formats: FLEEx

```
<paragraphs>
<paragraph guid="a9ca0c5a-1a26-4acc-b8fe-acbdda1acd72">
  <phrases>
    <phrase guid="9461ae9c-7290-477e-8205-b2d67f5cc90a">
      <item type="segnum" lang="fr">1</item>
      <words>
        <word guid="a95fa120-dc66-4d88-b53f-8a6c5ede925e">
          <item type="txt" lang="ru">К1к1ирда1лай</item>
          <morphemes>
            <morph type="stem" guid="d7f713e8-e8cf-11d3-9764-00c04f186933">
              <item type="txt" lang="ru">к1к1ирда</item>
              <item type="cf" lang="ru">к1к1ирди</item>
              <item type="variantTypes" lang="en" />
              <item type="gls" lang="en">Karata_people</item>
              <item type="glsAppend" lang="en">.obl</item>
            </morph>
            <morph type="suffix" guid="d7f713dd-e8cf-11d3-9764-00c04f186933">
              <item type="txt" lang="ru">-лай</item>
              <item type="cf" lang="ru">-лай</item>
              <item type="gls" lang="en">gen.pl</item>
            </morph>
          </morphemes>
        </word>
      </words>
    </phrase>
  </phrases>
</paragraph>
</paragraphs>
```

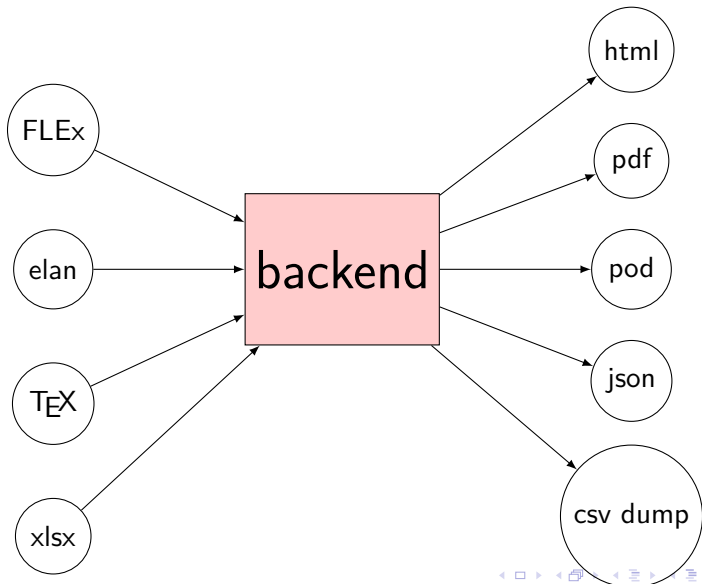
 ingestion via cldflex python library by Florian Matter



Input formats: tex

```
\ex a bēth thōk lē ka bēre\\  
\gll a bēth thōk lē ka bēre\\  
\textsc{1sg} cut tree \textsc{def} with axe\\  
\glt `I cut the tree with an axe.' (P67 K:2)
```

👉 possibly ingestion via `cldf-linglit` python library by Robert Forkel





- Cross-Linguistic Data Format

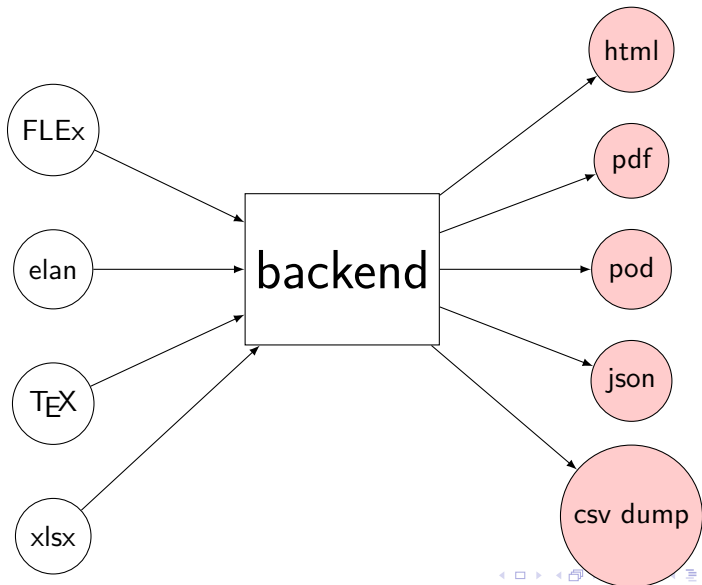
 - <https://clldf.clld.org>

- based on CSV

- standard in linguistic typology (grambank, dictionaria)

- cf. Forkel (this workshop)

	A	B	C	D	E
1	a↔beth↔thok↔le↔ka↔berε	1SG↔cut↔tree↔DEF↔with↔axe	'I cut the tree with an axe.'	P67 K:2	sher1258
2					
3					
4					
5					





Output formats: HTML

day passed|

[Clear search](#)

142 results
found in 17ms

Page size

5 10 25

Sorting

Filters

Language
iso6393

- aaz 3
- ayl 1
- chw 1

Mijał dzień za dniem.
pass.3SG.PST day.SG.NOM after day.SG.INS

Day after day passed. ()



Citation: docekal:wagiel:ed:21

 cf. Krämer & Nordhoff, this workshop



P. Epps, Isabel Salustiano, Jovino Monteiro & Pedro Pires Dias

- (11) *Yup m'é' sój d'öb k'ët yó' mah, tihàn tih íd-íh.*
yup m'é? sój d'öb-k'ët-yó?=mah, tih-àn tih ?id-ih.
 that carajuru brilliant descend-stand-SEQ=REP 3SG-OBJ 3SG speak-DECL
 'Standing there looking down, brilliant with *carajuru*, it's said, he spoke to her.'
 'Ficando lá olhando para ela, brilhante com carajuru, dizem, ele falou para ela.'
- (12) *Yít páh, "Hòp àmàn àh kāk w'öb péét, d'ó'óy àm páh?" nóoy mah.*
yít páh, hòp ?ám-àn ?āh kāk-w'öb-pé-ét, d'ó?-óy
 thus PROX.CNTR fish 2SG-OBL 1SG pull-set-go.upstream-DECL take-DYNM
?ám páh? nó-óy=mah.
 2SG PROX.CNTR say-DYNM=REP
 'And then, "Where I went upstream catching fish and setting them out for you; have you taken them?" he said, it's said.'
 'Aí, "Lá onde fui rio acima, pescando e deixando peixe, você pegou?" ele falou dizem.'



Output formats: PDF (community version)

wot kenambun kel-on go woh ne daman ko wot kol-on iyen adep e engg-un wot kenambun kel-on go woh ne daman ko wot kol-on iyen adep e engg-un enggan-u kedo aleng yanop kolo ambip wun-u-den aleng yanop kedo boma-n-u wen-e ambip adi nangg-e moyon ne daman =an engg-e olok wa-ngga-d-un engg-e tio-d-un got yanam angg-un=e nak-on ombet yu kaduk Awin ambip won-o-den odo mon-on mon-on go men-e keleg-e keleg-e tem-on keleg-e kem-on ga nowan wom to yu bet mo =on ma baat ode? ah baat odo yo ekune (a)dep an =o e om ko , baat Awin ambip won-o-den oyip mon-ok get =a om ko an-an-up =an o enggan-u engg-ain go ton a (go) b-e men-e bat boma-n-e b-e men-e kem-en odo wen-e om alep b-e mengga-mb-e kemo-d-on =e wen-e ekuni (a)dep kol-o-den o kole eh [ke] okun-o-den o , kelo-n-e wen-e nemengga-d-on

"The moon has become strong (more bright) so, oh maybe my brother has become the moon", she said. "The moon has become strong (more bright) so, oh maybe my brother has become the moon", she said. She said so and she went back home crying. She went away crying and arrived at home, she said "Oh, it is my younger brother" she said and she was missing him. She said that and stayed and true She slept and the next morning (next day) her husband who had gone to Awin region came back. He came and (he) looked around. He looked around but (there was) nobody. She was alone inside (the house). "But where is (my) brother-in-law?" "Oh, the brother-in-law" "it is like this" (, she said.) The sago, brother-in-law in Awin region he went so we will eat the sago when he comes back, she said and "I said like that and the fish and other things, that he usu-



Output formats: JSON-LD

```
▼ https://imtvault.org/content/static/ligt-0.2.ttl#hasWords:
  ▼ 0:
    @id: "https://imtvault.org/langsci316-6e388cec49_wt"
    ▼ @type:
      0: "https://purl.org/liodi/ligt#WordTier"
    ▼ https://imtvault.org/content/static/ligt-0.2.ttl#item:
      ▼ 0:
        @id: "._:langsci316-6e388cec49_0"
        ▼ https://imtvault.org/content/static/ligt-0.2.ttl#Word:
          ▼ 0:
            @language: "und"
            @value: "Mijał"
          ▼ 1:
            @language: "en-x-lgr"
            @value: "pass.3SG.PST"
          ▼ https://imtvault.org/content/static/ligt-0.2.ttl#nextWord:
            ▼ 0:
              @id: "._:langsci316-6e388cec49_1"
            ▼ 1:
              @id: "._:langsci316-6e388cec49_1"
              ▼ https://imtvault.org/content/static/ligt-0.2.ttl#Word:
                ▼ 0:
                  @language: "und"
                  @value: "dzień"
                ▼ 1:
                  @language: "en-x-lgr"
                  @value: "day.SG.NOM"
                ▼ https://imtvault.org/content/static/ligt-0.2.ttl#nextWord:
                  ▼ 0:
                    @id: "._:langsci316-6e388cec49_2"
                  ▼ 1:
                    @id: "._:langsci316-6e388cec49_2"
                ▼ https://imtvault.org/content/static/ligt-0.2.ttl#Word:
```



Output formats: CSV dump

- ↷ all sentences from all collections
- ↷ good for NLP use
- ↷ already done for IMTVault (Krämer & Nordhoff, this workshop)
- ↷ cf. Shu Okabe's presentation



Size

- ↗ aim: 5k words per collection
- ↗ 1.5 collections per region per year
- ↗ 5 regions, 3 years
- ↗ production of $\sim 40k$ word-sized tokens a year



As of today

40+ expressions of interest





Subsequent use

- ↪ automatically generated dictionaries
- ↪ typological research
- ↪ machine learning (“artificial intelligence”) for smaller languages
- ↪ (further uses presented and discussed during this workshop)



Future development

- ↪ 1-2 pilot collections per region to be published in 2024
- ↪ 2nd (or 3rd) generation effect > many more collections to be published until the end of the project
- ↪ Longevity through Language Science Press
 - ↪ OTC as a new series
- ↪ Integration into ELDP's training sessions for new grantees to assure high quality structured input



This workshop

- ♫ Christian Döhler (BBAW-OTC)
- ♫ Sebastian Nordhoff (BBAW-OTC)
- ♫ Mandana Seyfeddinipur (BBAW-ELAR)
- ♫ Kelsey Neely (BBAW-ELDP)
- ♫ Shu Okabe (Paris Saclay)
- ♫ Tom Liu (University Washington)
- ♫ Michael Ginn (University of Colorado)
- ♫ Florian Matter (Oregon)
- ♫ Hugh Paterson III (University of North Texas)
- ♫ Daniel Werning (BBAW-Wortschatz der ägyptischen Sprache)
- ♫ Thomas Krämer (GESIS – Leibniz-Institut für Sozialwissenschaften)
- ♫ Max Ionov (Universität zu Köln)
- ♫ Robert Forkel (MPI-EVA)



Thanks

- ↪ <http://opentextcollections.org>
- ↪ <https://fedihum.org/@otc>
- ↪ <https://twitter.com/OpenTextColl>
- ↪ facebook page: Open Text Collections



open
text
collections