

# Zwischen Forschungsdaten und Buchveröffentlichung

## Die Plattform Open Text Collections

Sebastian Nordhoff, Mandana Seyfeddinipur & Christian Döhler

Berlin-Brandenburgische Akademie der Wissenschaften

2023-09-28 Open-Access-Tage-Berlin

- Veröffentlichungsformen in der Sprachwissenschaft
  - Artikel
  - Bücher
  - Wörterbücher
  - Textsammlungen
  - (Multimedia)

# Open Access in den Sprachwissenschaften

- Generell gutes Bewusstsein für OA in der Disziplin
- Viele Sprachen werden in ökonomisch weniger wohlhabenden Weltregionen gesprochen
- Zugang ist für Sprecher und Forscher dort besonders wichtig.
- Artikel
  - [oaling.wordpress.com](http://oaling.wordpress.com)
  - Glossa (Diamond OA)
  - Zeitschrift für Sprachwissenschaft (Diamond OA)
- Bücher
  - Language Science Press (Diamond OA)
- Wörterbücher
  - Dictionaria
- Textsammlungen
  - ???

# Das Projekt *Open Text Collections*

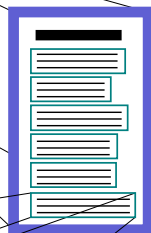
- Start letzte Woche
- DFG-Förderung 3 Jahre
- Danach konsortiale Finanzierung via Language Science Press
- innovativ, selektiv, international
- 5 Regional Boards:
  - Africa, Eurasia, Caucasus, Papunesia, South America

# Was ist eine Textsammlung?

**Sammlung**



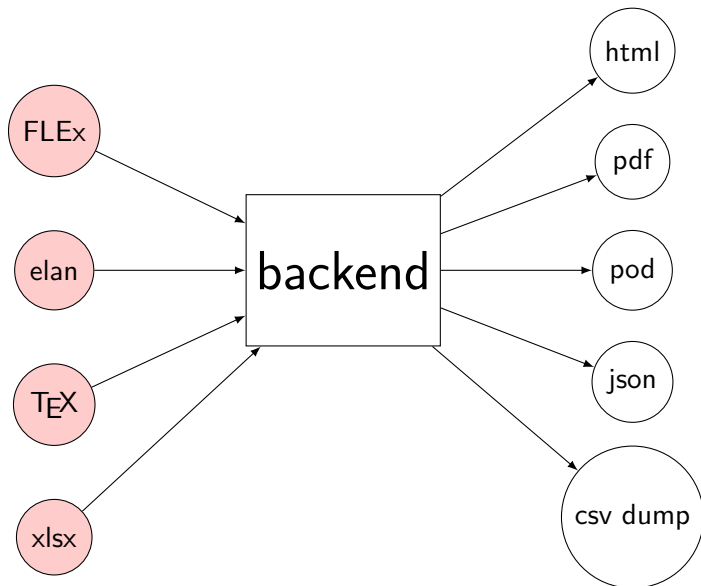
**Text**



**Satz**

Tisch-lein    deck    dich!  
table-DIM    set.IMP    2SG.ACC

'Table, set yourself!' ←



# Inputformate: ELAN

The screenshot displays the ELAN software interface with a timeline at the top showing time markers 00:02:07.500 and 00:02:08.000. Below the timeline, the sentence 'tór-tɕu-dze mè-òŋ-ge làp-ti' is segmented into individual words. A second row shows the morphological breakdown of each word. A third row shows the corresponding morpheme labels. A fourth row shows the glosses for each morpheme. A fifth row shows the morpheme classes and their corresponding glosses. A sixth row shows the number of occurrences for each morpheme class.

tór	tɕu	dze	mè	òŋ	ge	làp	ti
tór	tɕu	tɕe	mè	òŋ	-ke	làp	-di
lose	CAUS	INF	COP.NE	come	PRES	speak	IPFV
2	1	2					1

# Inputformat: FLEx

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   
Eng My Green Mat

Info Baseline Gloss Analyze Tagging Print View Text Chart

1.1 **Word** pus  yalola  nihimbilira .  
**Word Gloss** green  my mat I see  
**Word Cat.** mod  N 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

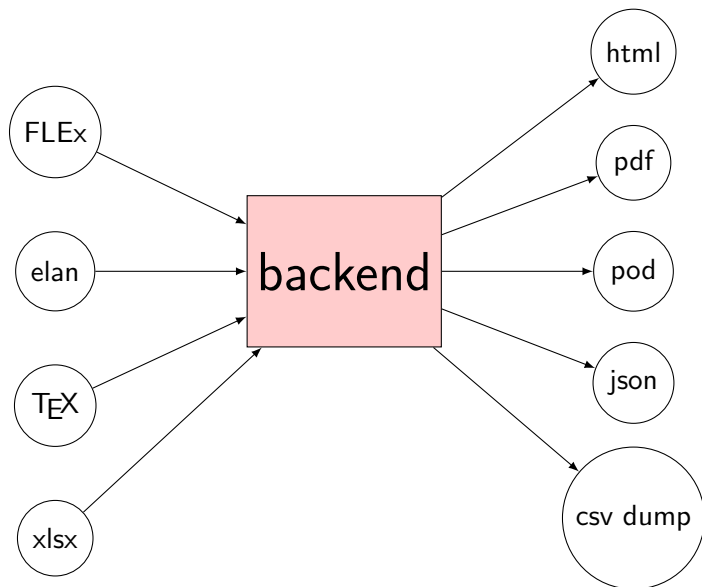
Lexicon  
Texts & Words  
Grammar  
Notebook  
Lists



## Inputformat: tex

```
\ex a beth thok le ka berε\\  
\gll a beth thok le ka berε\\  
\textsc{1sg} cut tree \textsc{def} with axe\\  
\glt `I cut the tree with an axe.' (P67 K:2)
```

# Backendformat

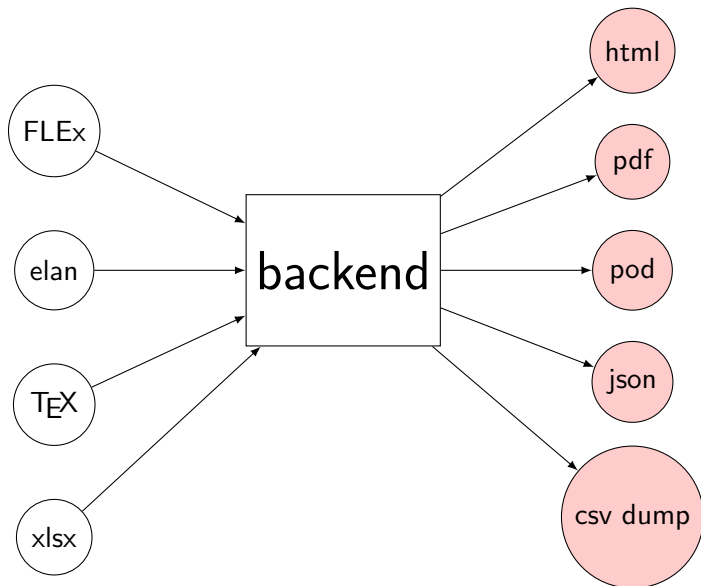


# Backendformat

- Cross-Linguistic Data Format
  - <https://cldf.clld.org>
- CSV-basiert
- sich abzeichnender Quasistandard in der Sprachypology

	A	B	C	D	E
1	aI→bethI→thokI→lcI→ka→berε	1SGI→cutI→treeI→DEFI→withI→axe	'I cut the tree with an axe.'	P67 K:2	sher1258
2					
3					
4					
5					

# Outputformate



# Outputformat: 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

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

- (11) *Yup m'é' sój d'öb k'ët yó' mah, tihàn tih ídíh.*  
*yup m'é? sój d'öb-k'ët-yó?=mah, tih-àn tih ?íd-íh.*  
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 ám páh?" nóóy 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.'

# Outputformate: JSON-LD

▼ <a href="https://imtvault.org/content/static/ligt-0.2.ttl#hasWords">https://imtvault.org/content/static/ligt-0.2.ttl#hasWords</a>	
▼ 0:	
@id:	"https://imtvault.org/langsci316-6e388cec49_wt"
▼ @type:	
0:	"https://purl.org/liodi/ligt#WordTier"
▼ <a href="https://imtvault.org/content/static/ligt-0.2.ttl#item">https://imtvault.org/content/static/ligt-0.2.ttl#item</a>	
▼ 0:	
@id:	"_:langsci316-6e388cec49_0"
▼ <a href="https://imtvault.org/content/static/ligt-0.2.ttl#Word">https://imtvault.org/content/static/ligt-0.2.ttl#Word</a>	
▼ 0:	
@language:	"und"
@value:	"Mijał"
▼ 1:	
@language:	"en-x-igr"
@value:	"pass.3SG.PST"
▼ <a href="https://imtvault.org/content/static/ligt-0.2.ttl#nextWord">https://imtvault.org/content/static/ligt-0.2.ttl#nextWord</a>	
▼ 0:	
@id:	"_:langsci316-6e388cec49_1"
▼ 1:	
@id:	"_:langsci316-6e388cec49_1"
▼ <a href="https://imtvault.org/content/static/ligt-0.2.ttl#Word">https://imtvault.org/content/static/ligt-0.2.ttl#Word</a>	
▼ 0:	
@language:	"und"
@value:	"dzierł"
▼ 1:	
@language:	"en-x-igr"
@value:	"day.SG.NOM"
▼ <a href="https://imtvault.org/content/static/ligt-0.2.ttl#nextWord">https://imtvault.org/content/static/ligt-0.2.ttl#nextWord</a>	
▼ 0:	
@id:	"_:langsci316-6e388cec49_2"
▼ 2:	
@id:	"_:langsci316-6e388cec49_2"
▼ <a href="https://imtvault.org/content/static/ligt-0.2.ttl#Word">https://imtvault.org/content/static/ligt-0.2.ttl#Word</a>	
▼ 0:	
@language:	"und"

# Textsammlungen als Hybrid zwischen Buch und Forschungsdaten

- Aus der Makroperspektive sind Textsammlungen Bücher
  - Autor\*in
  - Titel
  - Auflage
  - ISBN
  - Einband, ...
- Aus der Mikroperspektive handelt es sich um strukturierte Forschungsdaten
  - standardisierte Darstellung
  - Beziehung zwischen Elementen
  - Metadaten
- Ist die Bibliothek zuständig weil OA oder das Rechenzentrum weil Forschungsdatenmanagement?
- Was sind mögliche Finanzierungsmodelle?



- automatisch generierte/extrahierte Wörterbücher
- typologische Forschung
- maschinelles Lernen („künstliche Intelligenz“) für kleinere Sprachen

- OA bedingt den Export angelsächsischer/westlicher Konzepte
  - Eigentum
  - geistiges Eigentum
  - Rechtsstreit
  - verbrieftete Rechteabtretung
  - regional große Unterschiede
    - in Australien „gehören“ Sprachen den Sprechern
    - in Nordamerika große Vorbehalte gegen nicht-indigene Forschende
    - in den anderen Regionen weniger problematisch

# Publikationsanfragen Stand heute



Danke

<https://www.opentextcollections.org>

<https://fedihum.org/@otc>

<https://twitter.com/OpenTextColl>

