

# Corpora & Assistive Technology

The [Language Technology](#) group has expertise in handling various types of corpora. We are building tailor-made applications to explore large and structurally complex collections of language data. In particular, we are competent in:

- The design of databases to hold application-relevant data
- Generating interactive visualizations
- Efficiently querying large data collections (in particular corpora)
- Anonymisation of large data sets
- Data crawling/scraping and processing of web sources, batch download of documents
- Data extraction and conversion

## Examples of our work

Swissdox@LiRI – web based service for extraction of subcorpora from the Swiss media database  
Swissdox

Swissdox@LiRI    Start Projects Corpus query Retrieved datasets    Commercial Profile Test    Logout

### Corpus query

Languages \*

German  French

Source \*

20 Minuten (ZWA)

Date ranges

2023-01-01 – 2023-01-31

Document type \*

Select document types

Content keywords  Advanced content filter

Separate multiple entries by comma (e.g. Maske, COVID\*, ...). Search is case-sensitive and results will include articles with at least one of the keywords (OR operator). See more details in [Users' guide](#)

\* no filtering is applied if no option is selected

Reset filters  Next

<https://swissdox.linguistik.uzh.ch/>

VIAN – web application for multimodal corpora; comprises of corpus querying interface, multimodal corpus viewer, video and audio player and timeline with time-aligned text and annotations

VIAN

Document: AKAW1

eins zwei drei vier fünf sechs genau ja und die oben links

Frame: 3885 Time: 02:35/38.32

	00:00	00:01	00:02	00:03	00:04	00:05	00:06	00:07	00:08	00:09	00:10	00:11	00:12	00:13	00:14	00:15	00:16	00:17	00:18	00:19	00:20			
speaker 1 sentence	eh auf der eins zwei drei vier fünf sechs genau ja und die oben links heißt noch also das ist so ehm eine relativ komplizierte figur ehm es ist nicht v																							
Annotation	ADV	ADP	DET	NOM	NOM	NOM	NOM	NOM	ADV	CCONJ	DEP	PRON	AUX	ADV	PRON	DET	ADV	ADJ	NOUN	X	PRON	AUX		
Lemmas	eh	auf	der	einer	zwei	drei	vier	fünf	sechs	genau	ja	und	die	oben	links	ist	nicht	ein	komplizierte	figur	ehm	es	ist	nicht
speaker 1 right hand																								
speaker 1 left hand																								
speaker 2 sentence	auch sechs Figuren																							
Annotation	ADV	NUM	VERB																					
Lemmas	auch	sechs	Figuren																					
speaker 2 right hand	PG	OG																						
speaker 2 left hand																								

Query

```
[{"comment": "word \"hund\" in annotation layer word by SPK1", "nodes": [{"name": "et1", "layer": "tokens"}, {"filter": [{"AND": [{"attribute": "form", "values": "Noun"}]}]}], "Load example query: [1 2 3 4 5 6 7]"}]
```

Submit

80:63:27.4 hund AKAW1  
80:64:09.2 eh das sieht ziemlich nach einem hund aus AKAW1  
80:65:26.4 ein hund sein der am boden sitzt und der hat so ein AKAW1  
bisschen wie einen hund oder oder ein hund könnte es auch sein also ehm AKAW1  
also falls das ein hund ist hier AKAW1  
eine art hund sein oder so AKAW1  
ein hund sein der eigentlich also AKAW1  
gesicht also es könnte auch ein hund sein AKAW2  
dann als nächstes das ist definitiv ein hund der ehm ja so AKAW2  
da steht und ein bisschen nach unten wieder ein hund AKAW2  
hund aber er hat nicht wirklich einen kopf sonern AKAW2  
hund AKAW1  
wieder ein hund AKAW1  
der hund mit dem dreieck AKAW1

## CoLiCaSlav – web corpus application used as an empirical basis for teaching and studying the principle categories and concepts relevant for the Slavic languages

CoLiCaSlav

Home Search Categories References Sources Version 2

Language: BMSM Category: Select option Glossing: PL

Lemma: Select option Lemma (english): Select option Part of Speech (Upo): ADJ

bms: Tamo bismo bili sami ▾  
bms: To neću raditi ja, to čete uraditi vi sami ▾  
bms: Svi stadoše gledati jedni u drugi ▾  
bms: Taj savetnik je malopre kod Patrijarških ribnjaka ubio Mišu Berlioza. ▾  
bms: Na crnom su tržištu za audije postizali cijenu do 8000 njemačkih maraka, a volkswagenje i mercedese prodavalji su za 3000 maraka. ▾  
bms: Kapacitet mini pivovara prilagođeva se potrebama naručitelja, tako da je moguća proizvodnja od 600 do 6000 litara dnevno.Omogućena je proizvodnja različitih vrsta piva: svježih i uzelenih, s vodom ili nām sadržajem ekstrakta, a manje ili više alkohola. ▾  
bms: U posljednjih desetak godina otkriveni su arheološki ostaci helenističkoga grada, a u novije vrijeme pronađene su crkve poznate iz kasnoantoličkih i srednjovjekovnih spomenika. ▾  
bms: Ostali koji su iste večeri bili prevezeni u bolnicu, pušteni su kući. ▾  
bms: On sami napisu što misle o sebi. ▾  
bms: Sindikati su nezadovoljni predlaženim programom. ▾

<https://lehrkorpus-slav.linguistik.uzh.ch/>

## Kollo – command line tool for extracting collocations from VERT formatted corpora

```
usage: kollo [OPTIONS] [-r RIGHT] [-s SPAN] [-w [W,UL,DL,WL,UL,DL,T,A]] [-o STOPWORDS] [-t TARGET] [-n NUMBER] [-x] [-g] [-dsv] [CVR]
    Input [queries]

Extract collocations from VERT formatted corpora

positional arguments:
  input                Input file path
  query               optional regex to search for (i.e. to appear in all collocation results)

optional arguments:
  -h, --help            show this help message and exit
  -l LEFT, --left LEFT  window to the left in tokens
  -r RIGHT, --right RIGHT window to the right in tokens
  -s SPAN, --span SPAN  XML span to use as window (e.g. 3 or p)
  -W [W,UL,DL,WL,UL,DL,T,A], --window [W,UL,DL,WL,UL,DL,T,A]  collocation metric
  -o STOPWORDS, --stopwords STOPWORDS  Path to file containing stopwords (one per line)
  -t TARGET, --target TARGET  Index of VERT column to be searched as node
  -n NUMBER, --number NUMBER  number of top results to return (-1 will return all)
  -x OUTPUT, --output OUTPUT  Comma-sep index/indices of VERT column to be calculated as collocations
  -c, --case-sensitive  Do case sensitive search
  -p, --preserve        Preserve original sequential order of tokens in bigram
  -dsv [CSV], --dsv [CSV]  output comma-separated values

(pypy3.8) danny@danny-VirtualBox:~/Downloads$ kollo -textberg.vert "AIts1" --left 3 --right 4 --output 2 --number 20 --stopwords stopwords.txt
Finding collocates: 10000 | 38001 [00:14:00.00, 5,358bytes/s]
Building n-grams(texts): 38001 | 10000 [00:00:00.00, 22993.92bytes/s]
Formatting collocates: 10000 | 2378854/2378854 [00:11:00.00, 28587.47collocate/s]
Scoring collocates: 10000 | 2378854/2378854 [00:11:00.00, 28587.47collocate/s]
in      dass   66556.7626
an      steile  38549.9460
erleiste  ih    37480.3232
erleiste  er    34831.3784
erleiste  gallen 34831.3784
gallan   s     31385.6687
s       @card#  30058.9912
in      sommer 28379.0542
setze   se    28119.2348
se       setze  28119.2348
ih       vor    25276.1742
Haben   erleiste 25479.0484
losen   erleiste 25344.4566
@card#  stende 24955.5654
sehr   setzt 23991.9669
man   sv[galile] 22813.9984
man   sv[galile] 22813.9984
s       )     21785.0925
in      wie    19944.3714
ein      steht 39781.8788
```

From:

<https://liri.linguistik.uzh.ch/wiki/> - LiRI Wiki

Permanent link:

<https://liri.linguistik.uzh.ch/wiki/langtech/corpora>

Last update: **2023/02/02 14:27**

