

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

The screenshot shows the 'Corpus query' interface of the Swissdox@LiRI service. The interface is divided into several sections:



- Navigation:** A green header bar contains 'Swissdox@LiRI', 'Start', 'Projects', 'Corpus query', and 'Retrieved datasets'. On the right, there is a 'Commercial Profile Test' dropdown and a 'Logout' button.
- Form Fields:**
 - Languages *:** A dropdown menu with 'German' and 'French' selected.
 - Source *:** A dropdown menu with '20 Minuten (ZWA)' selected.
 - Date ranges:** A date range selector showing '2023-01-01 – 2023-01-31'.
 - Document type *:** A dropdown menu with 'Select document types'.
 - Content keywords:** A text input field.
 - Advanced content filter:** A text input field.
- Help and Actions:**
 - A note: '* no filtering is applied if no option is selected'.
 - Buttons for 'Reset filters' and 'Next'.
 - Instructions: 'Separate multiple entries by comma (eg. 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](#)'.

<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 Home Player Logout

Document: AKAW1

⏪ ⏩ ⏴ ⏵ 3 1 11 18 5.3s 1x 1.8x 2x 3x V1 V2 A1 A2

Frame: 3885 Time: 02:35/08:32

	02:32	02:33	02:34	02:35	02:36	02:37	02:38	02:39	02:40	02:41	02:42	02:43	02:44	02:45	02:46	02:47	02:48
speaker 1 sentence	eh	auf	der	eins	zwei	drei	vier	fünf	sechs	genau	ja	und	die	oben	links		
Annotation	ADV	ADP	DET	NUM	NUM	NUM	NUM	NUM	NUM	ADV	CCONJ	PRON	ADJ	NUM	VERB	ADV	PRON
Lemma	eh	auf	der	einer	zwei	drei	vier	fünf	sechs	genau	ja	und	der	oben	links		
speaker 1 right hand																	
speaker 1 left hand										hup!							
speaker 2 sentence		auch	sechs	Eigen										ein	eine	Huhn	der
Annotation		ADV	NUM	VERB										DET	DET	NOUN	NOUN
Lemma		auch	sechs	Eigen										ein	eine	Huhn	der
speaker 2 right hand	pg																
speaker 2 left hand	cg																

Query:

```
{
  "comment": "word 'hund' in annotation layer word by SPK1",
  "nodes": [
    {
      "name": "t1",
      "layer": "tokens",
      "filter": [
        "AND": [
          "attribute": "form",
          "value": "hound"
        ]
      ]
    }
  ]
}
```

Subrik

Load example query: 1 2 3 4 5 6 7

00:03:07.4	hund	AKAW1
00:04:09.2	eh das sieht ziemlich nach einem hund aus	AKAW1
00:05:26.8	ein hund sein der am boden sitzt und der hat so ein bisschen wie einen hut	AKAW1
00:11:11.8	oder oder ein hund könnte es auch sein also eh	AKAW1
00:18:58.3	also falls das ein hund ist hier	AKAW1
00:24:07.5	eine art hund sein oder so	AKAW1
00:26:17.5	ein hund sein der eigentlich also	AKAW1
00:27:38.5	gesicht also es könnte auch ein hund sein	AKAW2
00:32:10.3	dann als nächstes das ist definitiv ein hund der ihm ja so da steht und ein bisschen nach unten	AKAW2
00:39:26.3	wieder ein hund	AKAW2
00:39:36.4	hund aber er hat nicht wirklich einen kopf sondern	AKAW2
00:45:19.8	hund	AKAW1
00:45:52.3	wieder ein hund	AKAW1
00:46:32.8	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: BZMS Select all Deselect all Category: Select option Select all Deselect all Glossing: PL

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

bzms: Tamo bismo bili sami

bzms: To neću raditi ja, to ćete uraditi vi sami

bzms: Svi stadoše gledati jedni u drugu

bzms: Taj savetnik je malopre kod Patnjariškijh ribnjaka ubio Mišu Berliozu.

bzms: Na omron su tržištu za audije postojali cijenu do 8000 njemačkih maraka, a volkswagene i mercedese prodavali su za 3000 maraka.

bzms: Kapacitet mini pivovara prilagođava se potrebama naručitelja, tako da je moguća proizvodnja od 600 do 6000 litara dnevno. Omogućena je proizvodnja različitih vrsta piva-
svjetlih tamnih s višim ili nižim sadržajem ekstrakta, s manje ili više alkohola.

bzms: U posljednjih desetak godina otkriveni su arheološki ostaci helenističkoga grada, a u novije vrijeme pronađene su crkve poznate iz kaznoantičkih srednjovjekovnih spomena.

bzms: Onaj koji su iste večeri bili prevezani u bolnicu, pušteni su kuć.

bzms: Oni sami napisu što misle o sebi.

bzms: Sindikati su nezadovoljni predloženim programom.

Kollo - command line tool for extracting collocations from VERT formatted corpora

```
usage: kollo [-h] [-i INPUT] [-r RIGHT] [-s STOPWORDS] [-t TARGET] [-n NUMBER] [-o OUTPUT] [-c] [-p] [-w] [-C]
Input [query]

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
  -i INPUT, --input INPUT  file path to use as input (e.g. a or g)
  -n NUMBER, --number NUMBER  Collocation metric
  -s STOPWORDS, --stopwords STOPWORDS  Path to file containing stopwords (one per line)
  -t TARGET, --target TARGET  Index of VERT column to be searched as node
  -o OUTPUT, --output OUTPUT  Number of top results to return (-1 will return all)
  -c, --case sensitive  Case-esp index/indices of VERT column to be calculated as collocations
  -p, --preserve        Preserve original sequential order of tokens in ngram
  -cw [CSV], --csv [CSV]  Output comma-separated values

[py3.8] kollo@slav-ubuntu:~/kollo$ kollo lehrkorpus.vert --left 3 --right 4 --output 20 --stopwords stopwords.txt
Finding collocations: 100% |#####| 552/532M [01:43:00.00, 5.35Mbytes/s]
Building match contexts: 100% |#####| 163748/163716M [01:12:00.00, 22966.92Mbytes/s]
Formatting collocations: 100% |#####| 1553248/1553716M [00:28:00.00, 87093.18Mbytes/s]
Scoring collocations: 100% |#####| 2379864/2378864 [00:11:00.00, 288671.47Collocates/s]

so         dass         68550.7820
an         stelle       35549.9490
er/es/ste  is           37480.5232
er/es/ste  eta         38829.0718
st.,       gallen      34813.1174
scardg    s.          32282.6667
in         sommer      28370.0542
sete      so          28319.2346
so         sets        23821.9017
so         wir         25870.7421
haben     er/es/ste   25470.0464
lassen   er/es/ste   25394.6566
scardg   stende     24955.5054
lebr     stell       23891.9889
auf      sette      22913.9994
man      er/es/ste   22182.3602
s.       l           21745.8925
so       wte        19944.3714
ste      stende     19761.0788
```

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**

