



Friedrich-Alexander-Universität
Philosophische Fakultät und
Fachbereich Theologie



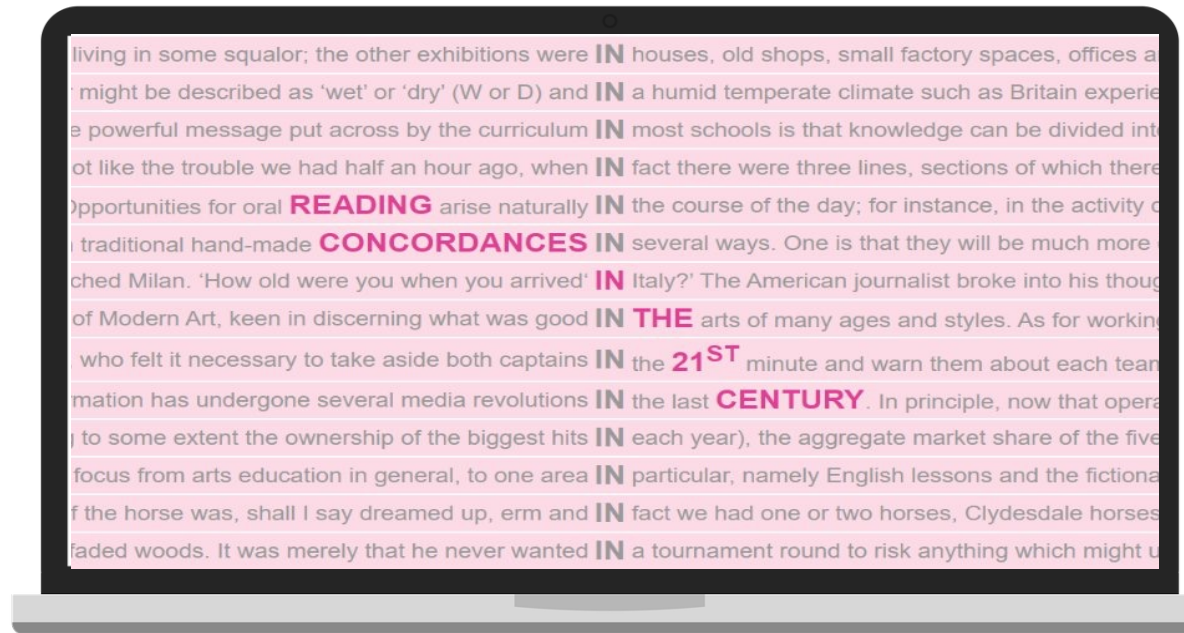
UNIVERSITY OF
BIRMINGHAM



Deutsche
Forschungsgemeinschaft



Arts and
Humanities
Research Council



5

Reading Concordances:

a training course in key corpus linguistics methodology

EESLI 2025 | 4 – 8 Aug 2025

@schtepf.bsky.social & @michamahlberg.bsky.social

Reading Concordances in the 21st Century (RC21) project team
Nathan Dykes • **Stephanie Evert** • **Michaela Mahlberg** • Alexander Piperski

Orientation

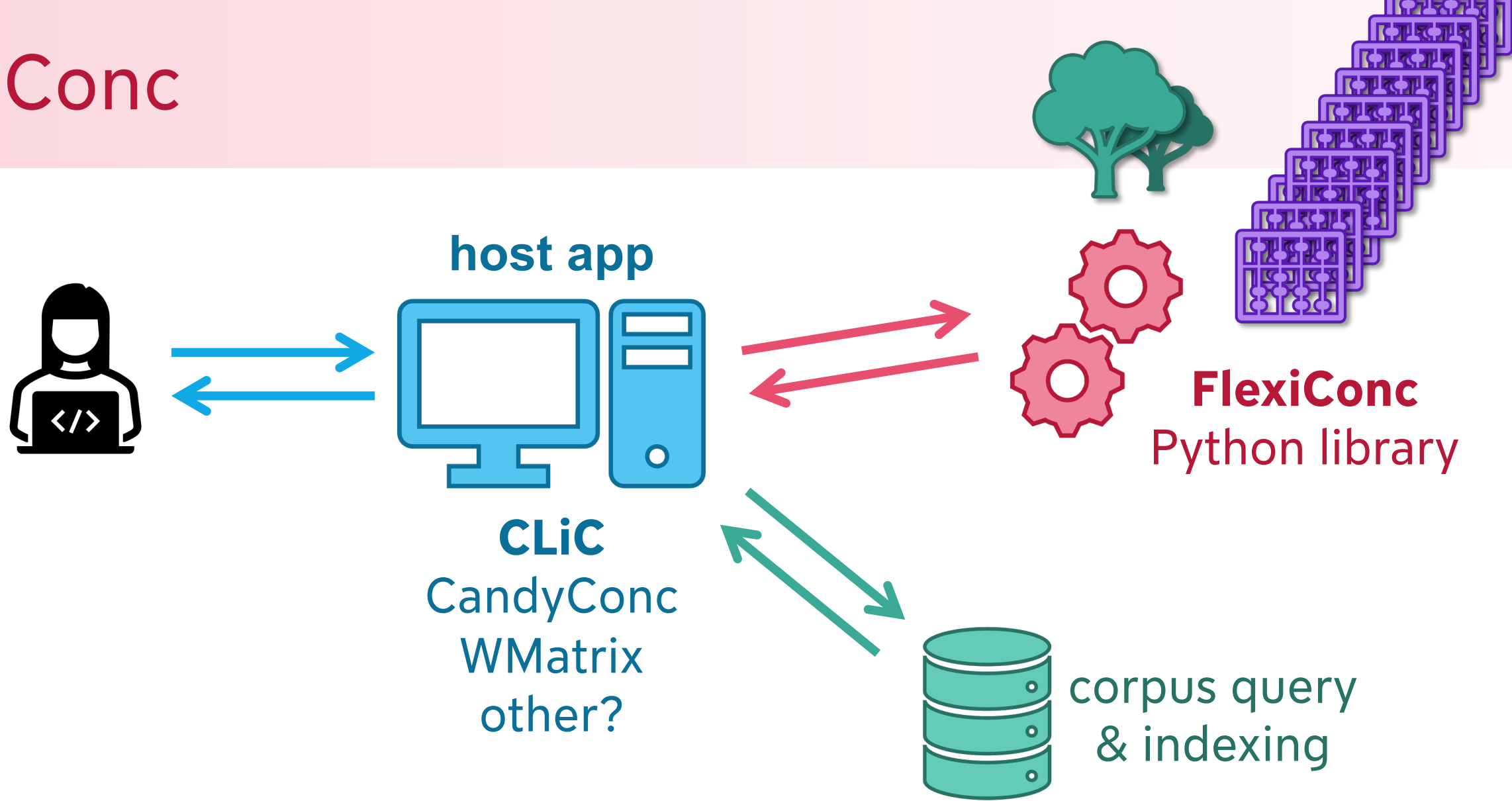
What we have learned so far ...

- Reading concordances with the help of software tools
- How to identify recurrent patterns and interpret them
- Strategies and algorithms for organising concordance lines (FlexiConc)
- The analysis tree as reproducible research documentation

What we're going to look at now ...

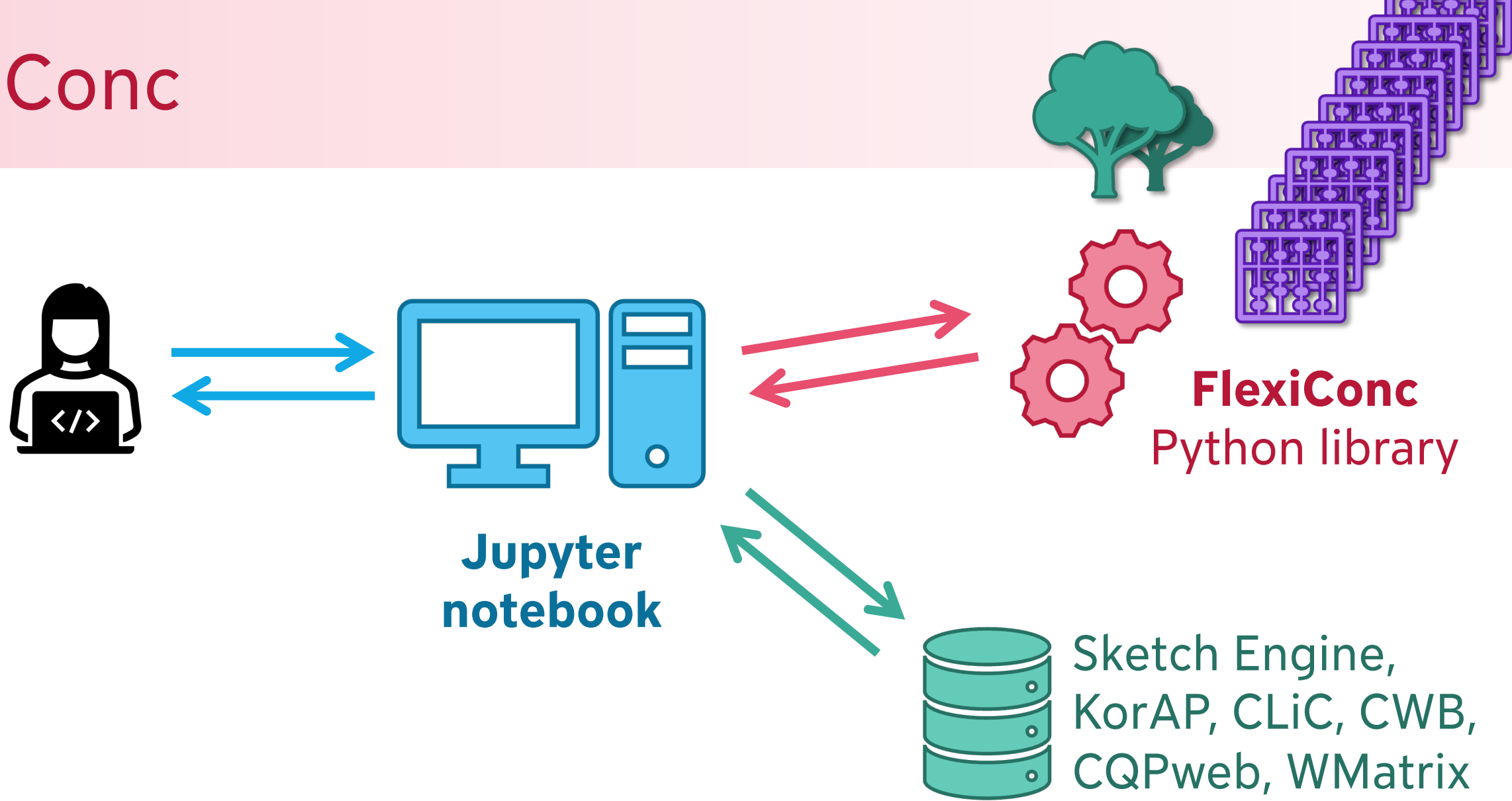
- Using FlexiConc from Jupyter notebooks
- How to analyse concordances in your own corpora

FlexiConc



<https://pypi.org/project/FlexiConc/>

FlexiConc



<https://pypi.org/project/FlexiConc/>

FlexiConc via Jupyter notebooks

Why FlexiConc? → benefit from algorithmic innovation!

- Use FlexiConc library from Python code in Jupyter notebook
- Combines (some) interactivity with custom programming, e.g. to carry out additional analyses not available in concordancer
- Connect with various indexing & search backend to access your own corpora there → FlexiConc has ready-made functions
- The Right Way: install Jupyter Lab + FlexiConc on your computer
- The easy way: run notebook in Google Colab ... [in a moment](#)

Installing FlexiConc (Windows / MacOS)

- install Anaconda Python or miniconda (anaconda.com/download)
- create a separate environment for FlexiConc
(`conda create -n FlexiConc python=3.13`)
- don't forget to activate it (`conda activate FlexiConc`)
- install PyICU from Anaconda (`conda install pyicu`)
- install JupyterLab (`conda install jupyterlab`)
- install FlexiConc and dependencies from PyPI with pip:
`pip install -U
"flexiconc[notebooks,web,ICU,partitioning,annotation]"`
- now start JupyterLab (`jupyter lab .`) ... or ask us for help

Installing FlexiConc (Linux)

- create virtualenv: `python3 -mvenv --upgrade-deps venv`
- activate virtualenv: `source venv/bin/activate`
- install ICU (e.g. `sudo apt install libicu-dev` on Ubuntu)
- install JupyterLab (`pip install jupyterlab`)
- install FlexiConc and dependencies from PyPI with pip:
`pip install -U
"flexiconc[notebooks,web,ICU,partitioning,annotation]"`
- start JupyterLab in your working directory: `jupyter lab .`
- ask us for help if it doesn't work ...

FlexiConc in Jupyter notebooks

- First steps: `flexicon_introduction.ipynb`



... or use in Google Colab

<https://colab.research.google.com/drive/1dvhZ4i0JplZDhrbH8BcGpeVOMN6RWXGa>

- Loading concordances into FlexiConc: `flexiconc_import.ipynb`



... or in Google Colab

<https://colab.research.google.com/drive/1xGl5tpcxv2-gVRxP5NjsPq4Ok66lwixc>

Import from CLiC

<https://clic-fiction.com/>

- Preparation: none

```
C = Concordance()  
C.retrieve_from_clic(query=["head"], corpora="dickens")
```

Import from CQPweb

<https://corpora.linguistik.uni-erlangen.de/cqpweb/>

Your query "water and sanitation" returned 669 matches in 424 different texts (in 409,134,520 words [1,956,223 texts]; frequency: 1.64 instances per million words)

[0.107 seconds - retrieved from cache]

Detailed output options

Formatting options

Choose operating system on which you will be working with the file: UNIX (incl. Mac OS X & iOS/Android) ▾

Print short handles or full values for text categories: full values ▾

Mark query results as <<< result >>>: No ▾

Size of context: 20 words each way ▾

Download both tagged and untagged version of your results: Yes ▾

Write information about table columns at the beginning of file: Yes - column headings ▾

Format of output - KWIC or line: KWIC ▾

Include sub-text region boundary markers: Yes ▾

Include corpus positions (required for re-import): Yes ▾

Include URL to context display: Yes ▾

Enter name for the downloaded file: CQPweb_WaterSanitation_ParLUK

Please tick the text metadata categories that you want to include in your download:

Method: Download text metadata ticked below ▾

Select from available text metadata:

- ☐ Agenda
- ☐ Chair?
- ☒ Date
- ☐ Month/Year
- ☒ Party
- ☐ Party Facts ID
- ☐ Speaker
- ☐ Speech number
- ☐ Terms
- ☒ Year

- ✓ Choose action...
- New query
 - Thin...
 - Frequency breakdown
 - Distribution
 - Dispersion
 - Sort
 - Collocations...
 - Download...**
 - Categorise...
 - Save current query result...

Go!

Import from CQPweb

<https://corpora.linguistik.uni-erlangen.de/cqpweb/>

- Obtain desired concordance in CQPweb and save it with download action (as shown on previous slide)
- Make sure to include all relevant metadata
- Save download file in same directory as Jupyter notebook

```
C = Concordance()  
C.load_from_cqpweb_export("CQPweb_WaterSanitation_Par1UK.txt")
```

Import from WMatrix

<https://ucrel-wmatrix7.lancaster.ac.uk/>

- **Tag Wizard:** Create your own corpus from text files in ZIP archive
- Annotated with POS, lemma, semantic concept in 9 languages
- Corpus ("folder") can be downloaded in SQLite format

Wmatrix7: Wmatrix multilingual tag wizard Wmatrix

You are logged in to Wmatrix7 as: demo1@esslli.2025 (1217)

[Tagging > Tag Wizard...]

[Folders > My folders | Details | Delete... | Archive... | Extract... | Library... | CrossTab... | Empty TRASH]

[Options > Switch to Simple Interface | Edit user options...]

[Help > Contents | Availability | Tagsets: POS & Semantic | USAS: Lexicon & MWEs & Context rules | Updates | Feedback]

[You are here > My folders]

Upload file

→

Part-of-speech tagging
Lemmatisation

→

Semantic tagging

→

MatrixDB indexing

→

Frequency lists
N-gram frequency lists
Collocation table

1. Choose language: English

2. Enter new folder name:

3. Click the button to select a file:
Choose file No file chosen

4. Upload now
Reset form

The Wmatrix tag wizard puts your corpus through the automatic language specific POS and Semantic analysis stages, indexes the results in the MatrixDB database, and produces frequency lists, n-gram frequency lists and a collocation table from your text file. For English text, this wizard runs the CLAWS tagger and USAS tagger. For other supported languages, the wizard runs the PyMUSAS tagger tagging pipeline. For English, please do not run large texts (e.g. a file with more than 1 million words, or a collection of files in a zip where any one file is larger than 1 million words) through the tag wizard. These are better run off-line and loaded into Wmatrix afterwards. Please get in touch with Paul to do this. For texts run through PyMUSAS, spaCy currently sets a default maximum length per file of 1 million characters.

File types:
Please view [Tutorial A](#) for details on how to convert your PDF, DOC(X) or RTF files to TXT format which is suitable for Wmatrix. Further [input format guidelines](#) are available for the English CLAWS/USAS taggers, including for example how to avoid problems with less-than and greater-than symbols in the input text.

One corpus per folder:
If you do not specify a folder, one will be created with a unique name. It is recommended that you use a **new folder for each corpus**. If your corpus consists of more than one file, then we recommend concatenating the files together first or using a zip file to load all the files together.

i

©2000-25 UCREL, Lancaster University.
For technical queries please contact Paul Rayson : p.rayson@lancaster.ac.uk

Import from WMatrix

<https://ucrel-wmatrix7.lancaster.ac.uk/>

- Use own corpus ("folder") or copy **ESSLI_Water_Par1UK** from library
- Main function: keyword analysis (for word, lemma, POS, concept)
- Note interesting keywords → concordance analysis in FlexiConc

Wmatrix7: Folder ESSLI_Water_Par1UK

You are logged in to Wmatrix7 as: demo1@essli.2025 (1217)

[**Tagging** > Tag Wizard...]


[**Folders** > My folders | Details | Delete... | Archive... | Extract... | Library... | CrossTab... | Empty TRASH]

[**Options** > Switch to Simple Interface | Edit user options...]

[**Help** > Contents | Availability | Tagsets: POS & Semantic | USAS: Lexicon & MWEs & Context rules | Updates | Feedback]

[You are here > My folders > ESSLI_Water_Par1UK]

| | Frequency list | Concordance | N-grams | Collocation | Keyness analysis |
|----------------|---|-------------|---------|-------------|---|
| Word | Word only (Sorted by: Frequency ; Word) | | 2 3 4 5 | Word | Key words compared to: <input type="text" value="British English 2021 (BE21)"/> <input type="button" value="Go"/> |
| Lemma | Lemma only (Sorted by: Frequency ; Lemma) | | | | Key lemmas compared to: <input type="text" value="British English 2021 (BE21)"/> <input type="button" value="Go"/> |
| Part of speech | POS only (Sorted by: Frequency ; POS) Word and POS (Sorted by: Frequency ; Word ; POS) | | | | Key POS compared to: <input type="text" value="British English 2021 (BE21)"/> <input type="button" value="Go"/> |
| Semantic | USAS Tag only (Sorted by: Frequency ; USAS tag) Word and USAS tag (Sorted by: Frequency ; Word ; USAS tag) | | | | Key concepts compared to: <input type="text" value="British English 2021 (BE21)"/> <input type="button" value="Go"/> |



Import from WMatrix

<https://ucrel-wmatrix7.lancaster.ac.uk/>

- Get a free WMatrix account (or use one of our demo accounts)
- Download complete annotated WMatrix corpus
- Then use query to obtain concordance for desired keyword

```
labour2005 = wmatrix.load(  
    corpus_name="LabourManifesto2005",  
    username="[USER]", password="[PASSWORD]",  
    db_filename="labour2005.db")  
C = labour2005.concordance_from_query(  
    r'[lemma="community" %c]')
```

Import from Sketch Engine

https://app.sketchengine.eu/#ca?corpname=user%2FSEvert%2Ftta

MANAGE CORPUS

TTA

Get more space

My account

My Sketch Engine

Settings

Local administration

Logout

CORPUS: TTA (English)

Trump Twitter Archive

Browse

View documents and folders, edit metadata

Delete

Remove corpus permanently

Stephanie Evert (ID: 80014)

SUBSCRIPTION & INVOICING

CHANGE PASSWORD

Username

SEvert

E-mail

stephanie.evert@fau.de

Account type

Multi-user account

Group account end date

April 8, 2026

Corpus storage used (words)

4,098,805 of 100,000,000 (4%)

Academic user

yes

Sketch Engine API key

78b4f7fd6c3e75e62469f535d88fcd69

Generate new key

Stephanie Evert

Account admins

Admins can add or pause users and change their storage space.

CLOSE

New corpus

Create new corpus

Import from Sketch Engine

sketchengine.eu/

TTA user/SEvert/tta created May 7, 2025 at 8:52:32 PM

Trump Twitter Archive

MANAGE CORPUS

MANAGE SUBCORPORA

COMPARE CORPORA

TEXT TYPE ANALYSIS

GENERAL INFO

Language: English

CORPUS DESCRIPTION & BIBLIOGRAPHY

TAGSET

WORD SKETCH GRAMMAR

TERM GRAMMAR

COUNTS

| | |
|-----------|-----------|
| Tokens | 1,359,510 |
| Words | 1,036,092 |
| Sentences | 377 |
| Documents | 56,570 |

TEXT TYPES

TEXT TYPE ANALYSIS

| | |
|--------------------------|--------|
| <doc> (2) | 13 |
| File ID , doc.id | 13 |
| File name , doc.filename | 13 |
| <text> (0) | 56,570 |
| <g> (0) | 27,609 |
| <s> (0) | 377 |

LEXICON SIZES

| | |
|-----------|--------|
| word? | 75,630 |
| tag | 63 |
| lempos? | 49,323 |
| pos | 9 |
| lemma | 46,084 |
| lempos_lc | 46,180 |
| lemma_lc | 42,597 |
| lc | 64,983 |

COMMON TAGS

| | |
|-------------|------|
| adjective | J.* |
| adverb | RB.? |
| conjunction | CC |
| determiner | DT |
| noun | N.* |
| numeral | CD |
| particle | RP |
| preposition | IN |
| pronoun | PP.? |
| verb | V.* |

All tags

Import from Sketch Engine

<https://app.sketchengine.eu/>

- Prerequisite: paid SkE account (free 30-day trials available)
- Generate API access token (as shown on previous slides)
- Note down full path to desired corpus (as shown on previous slides)

```
C = Concordance()  
C.retrieve_from_sketchengine(  
    query='[lc="fake"] [lc="news"]',  
    corpus="user/SEvert/tta", # insert your path here  
    api_key="[YOUR API KEY]")
```

Pro users: Import from CWB

- Install IMS Open Corpus Workbench from cwb.sourceforge.io
- Index your favourite corpus (or obtain a pre-indexed one)
- Python API: `pip install cwb-ccc`
- NB: will need complete development environment with C compiler

```
C = Concordance()
C.retrieve_from_cwb(
    corpus_name = "PARLSPEECH_UK",
    query = r'[pos="JJ.*"] [lemma="elephant"]',
    tokens_attrs = ["word", "pos", "lemma"],
    metadata_attrs = ["text_party", "text_year", "text_date"])
```

Wrapping up ...

- Concordance reading as link between quantitative and qualitative perspectives on corpora (“distant vs. close reading”)
- Many applications: literary stylistics, discourse analysis, language teaching, lexicography, query refinement, topic models, ...
- Aim: identify recurrent patterns at different levels → interpretation
- Computational algorithms support organising concordance lines
- Strategies: Selecting, Sorting, Ranking, Partitioning, Clustering
- Analysis tree as reproducible research documentation & template
- FlexiConc Python library as open-source software implementation

Farewell & materials

Thank you!

- Course homepage:
<https://www.dhss.phil.fau.eu/essli-2025-reading-concordances/>
- Project homepage:
<https://www.dhss.phil.fau.eu/research/reading-concordances/>
- Open course materials for reuse (soon):
<https://github.com/reading-concordances/teaching/>



Contributions welcome!

Please fill in our feedback form for the ESLLI course!

<https://forms.gle/mVndts8zHrcoWR2T9>



Home / Research / Current Projects / Reading Concordances in the 21st Century (RC21) /

RC21 Blogs

< Research

Current Projects

Reading Concordances in the 21st Century (RC21)

RC21 Blogs


RC21 Events

Publications

Talks

Doctorates


Concordance Reading × Association Measures



11. June 2025

Category: [RC21](#)

Meeting corpus users' needs



Concordance Reading × Association Measures Methods to Organize a Dataset of German Support Verb Constructions Author: Xinyao Lu, Friedrich-Alexander-Universität Erlangen-Nürnberg Published: 11 June 2025 Introduction How can we collect a dataset of German Support Verb Constructions (SVCs) fr...

Continue >

Meeting corpus users' needs Author: Yukio Tono (Tokyo University of Foreign Studies) Published: 17 January 2025 As language tools evolve in the digital age, having the right tools can make all the difference in how we interact with and analyze language. But what makes a tool truly effective...

Get in touch with us to contribute to our project blog!

- experience with FlexiConc
- concordance reading for your own research
- needs/ideas for algorithms

8 Aug 2025 | © RC21 Team

Reading concordances

21