Online and Desktop Tools for Querying a Language Corpus

CLARIN ERIC
15 September 2022
Organisers

This edition of the CLARIN Café is organized by

Martin Wynne, Senior Researcher in Corpus Linguistics and National Coordinator for CLARIN-UK, University of Oxford

CLARIN hosts is
Francesca Frontini (CLARIN BoD)

Technical support by
David Bordon

The event is recorded for further dissemination purposes.

Questions and comments? Put them in the chat box.
Schedule

14:00 - 14:10 Welcome and Introduction. Francesca Frontini, CLARIN

14:10 - 14:30 Tools for corpus analysis: a survey and guide to finding the right one. Martin Wynne, University of Oxford, UK

14:30 - 14:50 #LancsBox: a new-generation software package for the analysis of language data and corpora. Vaclav Brezina, Lancaster University, UK

14:50 - 15:10 AntConc: a corpus analysis toolkit for concordancing and text analysis. Laurence Anthony, Waseda University, Japan

15:10-15:15 break

15:15 - 15:35 Kontext: an online corpus query interface and its use in CLARIN. Michal Kren, Charles University, Prague

15:35 - 16:00 Discussion panel: what steps can we take to make corpus tools easier to find, use and connect together?
CLARIN 101

https://www.clarin.eu/content/clarin-in-a-nutshell
CLARIN ...

- is the *Common Language Resources and Technology Infrastructure*
- has the **ESFRI** ERIC status since 2012, Landmark since 2016
- provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
  - to **digital language data** (in written, spoken or multimodal form)
  - and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
  - through a **single sign-on** environment
- serves as an ecosystem for **knowledge sharing and training**
- is one of the European RIs in the SSH cluster (aka SCI)
- is an integral part of **the European Open Science Cloud**
  - See [clarin.eu/eosc](http://clarin.eu/eosc)
CLARIN today

- a distributed network of **70 centres**
- **22 members:** AT, BE, BG, CY, CZ, DE, DK, EE, FI, GR, HR, HU, IS, IT, LT, LV, NL, NO, PL, PT, SE, SI
- **2 observers:** UK, ZA
- **1 third party**
The Technical Infrastructure

Harvested Metadata

Metadata Search
Virtual Language Observatory (VLO)

Data Analysis
Language Resource Switchboard (LRS)

clarin.eu/fair
vlo.clarin.eu
switchboard.clarin.eu
The Knowledge Infrastructure

Knowledge Centres
Tour de CLARIN
Annual Conference
Video Channels

Digital Humanities Course Registry
Teaching
Funding
Best-Practice Papers

https://www.clarin.eu/content/knowledge-infrastructure
Resource Families

• What is CRF?
  – user-friendly overviews per data type of the available language resources in CLARIN
  – https://www.clarin.eu/resource-families
  – 13 corpora families, 5 families of lexical resources, and 4 tool families
  – the most important metadata and brief descriptions as well as links to download pages and concordancers

• Funding for small projects that can help extending the scope of the initiative:
  https://www.clarin.eu/content/clarin-resource-families-project-funding
The Café
Getting involved in CLARIN

Join our NewsFlash
https://www.clarin.eu/content/newsflash

Check out our events
https://www.clarin.eu/events

Open calls
https://www.clarin.eu/content/funding-opportunities
See you at the CLARIN Conference

Register for Virtual Attendance:


Next events:

Stay tuned: https://www.clarin.eu/content/

Share your #clarincafe impressions with @CLARINER
Tools for corpus analysis: a survey and guide to finding the right one

Martin Wynne
martin.wynne@ling-phil.ox.ac.uk

Faculty of Linguistics, Philology and Phonetics, University of Oxford
National Coordinator, CLARIN-UK

https://llds.ling-phil.ox.ac.uk/

CLARIN Café:
Online and Desktop Tools for Querying a Language Corpus
Thursday 15th September 2022

Arts and humanities infrastructure enabling knowledge with impact

Tao Chang
Head of Infrastructure, Arts and Humanities Research Council

1 September 2021

Arts and humanities infrastructures are vital for driving innovation in heritage science, and ensuring the creative industries thrive.

https://llds.ling-phil.ox.ac.uk/
Data & Tools

- CLIC
  An online corpus for exploring the language of Dickens

- GATE
  General Architecture for Text Engineering

Latest News

- Free corpus linguistics events: 20-23 June
  14 June 2022

- Horizon Europe guarantee notice and guidance
  Martin Wynne, 26 May 2022

Events

- Online and Desktop Tools for Querying a Language Corpus
  Thursday 15 September, 13:00 to 15:15 (BST), Online

- CLARIN Annual Conference 2022
  Monday 10 October, Prague, Czechia

Tweets from @clarin_uk

- CLARIN UK Retweeted
  CLARIN ERC @CLARINERC · 6h
  You are still on time to register for tomorrow's Online and Desktop Tools for Querying Corpus, starting at 14:00
  More info and registration here: clarin-er.eu @antlabp @Korpus_cz @vaclavrezina @f_frontini

https://www.clarin.ac.uk/
CLARIN Resource Families

Corpora
- Computer-mediated communication corpora
- Corpora of academic texts
- Historical corpora
- L2 learner corpora
- Legal corpora
- Literary corpora
- Manually annotated corpora
- Multimodal corpora
- Newspaper corpora
- Parallel corpora
- Parliamentary corpora
- Reference corpora
- Spoken corpora

Lexical Resources
- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

Tools
- Normalization
- Named entity recognition
- Part-of-speech tagging and lemmatization
- Tools for sentiment analysis

Coming soon:
- Online Corpus Tools
- Desktop Corpus Tools

https://www.clarin.eu/resource-families
Online and desktop tools for corpus analysis

“Corpus, concordance, collocation”
Please contribute by adding comments to the spreadsheet at:

<table>
<thead>
<tr>
<th>Title</th>
<th>Link</th>
<th>Country (where website is based, if relevant)</th>
<th>Free / not free (money)</th>
<th>Registration required - do users need to register to use the tool?</th>
<th>Shibboleth authentication? (Can users log in with their own institutional single-sign on?)</th>
<th>Can you upload your own corpus?</th>
<th>Support (may have support as part of a project but time limited so not sustainable), say YES if eg an email address given</th>
<th>Citation (how to cite the application, if this information is available)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>COSMAS 2</td>
<td><a href="https://cosmas2.ids-mannheim.de/cosmas2-web">https://cosmas2.ids-mannheim.de/cosmas2-web</a></td>
<td>Germany</td>
<td>Free</td>
<td>No (certain authorities granted to users with email in the .ac.uk domain)</td>
<td>Yes</td>
<td>Yes</td>
<td>Online help system, bug reporting system, email address for queries</td>
<td>Yes - no indication of how obtain credentials in order to authenticate</td>
<td>the German reference corpus - DeRaKo, historical corpora and others.</td>
</tr>
<tr>
<td>COSMAS II</td>
<td><a href="https://cosmas2.ids-mannheim.de/cosmas2-web">https://cosmas2.ids-mannheim.de/cosmas2-web</a></td>
<td>Germany</td>
<td>Free</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Will be superseded by Ziggurat!</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CQPweb (Lancaster)</td>
<td><a href="https://cqpweb.lancs.ac.uk/">https://cqpweb.lancs.ac.uk/</a></td>
<td>UK</td>
<td>Free</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td>Numerous South-Asian languages</td>
</tr>
<tr>
<td>CQPweb (Beijing Foreign Studies University)</td>
<td>CQPweb</td>
<td>China</td>
<td>Free</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td>NoScriptEngine interface to the Chinese Language Resources and Infrastructure (CRI)</td>
</tr>
<tr>
<td>Croatian National Corpus</td>
<td><a href="http://filip.ffng.hr/">http://filip.ffng.hr/</a></td>
<td>Croatia</td>
<td>Free</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td></td>
<td>The Croatian National Corpus</td>
</tr>
<tr>
<td>DWDS</td>
<td><a href="https://www.dwds.de/">https://www.dwds.de/</a></td>
<td>Germany</td>
<td>Free</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td></td>
<td>The DWDS is part of the Center for Digital Lexicography of the German Language (ZDL), funded by the Federal Ministry of Education and Research. It is based at the Berlin-Brandenburg Academy of Sciences</td>
</tr>
</tbody>
</table>

CLARIN corpus resources and tools

Corpora: at least 4130 - see VLO (https://vlo.clarin.eu/)!

Online interfaces:

- **Corpuscle**
- **Korp**
- **KonText**
- **NoSketch Engine**
- **D* (Diacollo demo)**
- **TEITOK**

Federated content search: https://contentsearch.clarin.eu/

Resource Families:

- 13 curated guides to different types of corpora and how to get them
- Coming soon: *Desktop corpus tools* and *Online corpus tools*
Diachronic collocations in a text collection: DiaCollo from the Deutsches Textarchiv
<table>
<thead>
<tr>
<th>Number</th>
<th>Page</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1580:8</td>
<td>Altar / daß daselbst warhaftig der Leib vnd Blut im Brot vnd Wein / werde mündlich geessen...</td>
</tr>
<tr>
<td>2</td>
<td>1580:10</td>
<td>... Brot vnd Wein / seinen warhaftigen Leib vnd Blut / zuessen vnd zu trinken darerche vnd gebe...</td>
</tr>
<tr>
<td>3</td>
<td>1580:18</td>
<td>... die unglaubige vnd vnwürdige / den Leib vnd Blut Christi mit Brot vnd Wein / im heiligen...</td>
</tr>
<tr>
<td>4</td>
<td>1580:19</td>
<td>... Wein / aber gar nicht der Leib vnd Blut Christi / gegenwertig dargereicht vnt empfangen werde /...</td>
</tr>
<tr>
<td>5</td>
<td>1580:31</td>
<td>... wir vns / ob der ware Leib vnd Blut Christi leiblich im Brot vnd Wein sey...</td>
</tr>
<tr>
<td>6</td>
<td>1580:33</td>
<td>... anders gewolt / dann daß Christi Leib vnd Blut warhaftig im Nachtmal mit Brot vnd Wein...</td>
</tr>
<tr>
<td>7</td>
<td>1580:35</td>
<td>... von einer wesentlichen Gegenwertigkeit des Leibs vnd Blut Christi / vnder gesallt des Brots vnd Weins...</td>
</tr>
<tr>
<td>8</td>
<td>1580:36</td>
<td>... Brots vnd Weins / in den Leib vnd Blut Christi (also daß kein Brot vnd Wein...</td>
</tr>
<tr>
<td>9</td>
<td>1580:36</td>
<td>... Brots vnd Weins / in den Leib vnd Blut Christi (also daß kein Brot vnd Wein...</td>
</tr>
<tr>
<td>10</td>
<td>1580:37</td>
<td>... vnd mit dem sichtbaren Wein / das Blut Christi dargereicht werde.</td>
</tr>
</tbody>
</table>

Information headings

- Name of the application
- Link (preferably to landing page with information about the tool and how to get it)
- Available for download (yes/no or give URL if different to column B)?
- Licence - according to the terms of which user licence can the application be used?
- Citation (how to cite the application, if this information is available)
- Platform-dependence (is it only available to work on specific platforms or operating systems?)
- Free / not free (money)
- Limited dataset - does it only work with the included corpus?
- Support - is user support available?
Summary

The questions we are trying to answer:
● As a user, how do I choose which application to use for corpus analysis?
● As a software engineer, how do I make it easier to users to find and use my software?
● As infrastructure builders, how do we make tools easier to deploy, more sustainable, and more able to interoperate with different datasets and tools?

Get involved!
● Tell us what more we can do to help answer these questions;
● Tell us what other questions we should be asking!

National contacts:
See https://www.clarin.eu/content/participating-consortia

Contact the workshop organisers here:
francesca.frontini@ilc.cnr.it
martin.wynne@ling-phil.ox.ac.uk

General contact:
clarin@clarin.eu
#LancsBox: a new-generation software package for the analysis of language data and corpora

Vaclav Brezina
Lancaster University
@vaclavbrezina
Corpus linguistics in 2022

- More popular than ever before.
- More interdisciplinary than ever before.
- More complex and sophisticated than ever before.
- More challenging than ever before.
Multidimensional data

- Formal vs. informal
- Speech vs. writing
- Newspapers vs. fiction vs. academic writing vs. blogpost vs. tweets...
- Young vs. old
- Male vs. female
- Historical vs. present
- L1 vs. L2 production
innovation in corpus linguistics

#LancsBox

@Lancaster University
Current version: #LancsBox v. 6: Corpora, Smart searches, Wizard, Automatic corpus building...

New version: #LancsBox X
Very large corpora (billions of words), XML, new UI
Corpus Linguistics: Method, Analysis, Interpretation

4.9 (95 reviews)

Get a practical introduction to the methodology of corpus linguistics for researchers in the social sciences and humanities.

Join course

70,092 enrolled on this course
Feedback cycle

Roberto Esposito
Follow 27 NOV

It is getting funny. Human brain reacts enthusiastically when eyes are placed in front of graphs, colours, maps and schemes. It gets kinda challenging to get out it. I like it, and I think it will be very useful!

Andrea Mosquera González
Follow 25 NOV

The graph feature of Lancsbox is very helpful and it was very fascinating to see how it works. Thank you!

Becky H
Follow 21 NOV

The graph feature is super helpful - I will be using this with my ESL students :)

Feedback cycle

1. User feedback
2. Development team
3. Bug fixes
4. Innovations
5. #LancsBox feature
Download #LancsBox:
http://corpora.lancs.ac.uk/lancsbox/
#LancsBox highlights
1. Smart searching

Smart searches

- NOUN
- PROPER NOUN
- VERB
- ADJECTIVE
- ADVERB
- MODAL
- CONNECTOR
- PRONOUN
- CONTRACTION
- PASSIVE
- COMPLEX NOUN PHRASE
- PAST TENSE
- NOMINALIZATIONS
- SPLIT INFINITIVE
- PRESENT TENSE
- PAST TENSE
- PLACE ADVERBIAL
- TIME ADVERBIAL etc...
2. CQL: Corpus Query Language

[word="goes" & headword="go" & pos="V.* "]

[pos="VB.*"] [][]{0,3} [pos="V.N"]
2. #LancsBox X: new UI
3. #LancsBox X: very large corpora
<table>
<thead>
<tr>
<th>File</th>
<th>Left</th>
<th>Node</th>
<th>Right</th>
<th>Text: da</th>
</tr>
</thead>
<tbody>
<tr>
<td>LordChan</td>
<td>everyone has access to healthcare. We will invest more in vaccines</td>
<td>and research, so that developing countries can benefit</td>
<td>2020-01-07</td>
<td></td>
</tr>
<tr>
<td>LordChan</td>
<td>The same is true for GAVI, as the efficacy of vaccines is massively reduced if the recipient is malnourished.</td>
<td></td>
<td></td>
<td>2020-01-07</td>
</tr>
<tr>
<td>LordChan</td>
<td>to spend large sums of money on girls’ education, vaccinations</td>
<td>or climate change, let us not just tick the box,</td>
<td>2020-01-07</td>
<td></td>
</tr>
<tr>
<td>LordChan</td>
<td>ment on their work in these areas. Education, climate and vaccinations are key building blocks for creating a world in which</td>
<td></td>
<td></td>
<td>2020-01-07</td>
</tr>
<tr>
<td>LordChan</td>
<td>it is also why we have run a vigorous flu vaccination programme to prevent people from getting into that</td>
<td></td>
<td></td>
<td>2020-01-09</td>
</tr>
<tr>
<td>LordChan</td>
<td>clinical trial into scurry, to proton beam therapy and mass vaccination</td>
<td>programmes. The Medicines and Medical Devices Bill</td>
<td>2020-01-09</td>
<td></td>
</tr>
<tr>
<td>CommCh</td>
<td>children in this country are not protected by the MMR vaccine when they start school? May we have an urgent debate</td>
<td></td>
<td></td>
<td>2020-01-09</td>
</tr>
<tr>
<td>CommCh</td>
<td>doing what they can to increase the take-up of measles vaccines to ensure that as many children as possible are covered</td>
<td></td>
<td></td>
<td>2020-01-09</td>
</tr>
<tr>
<td>CommCh</td>
<td>have a great responsibility to ensure that their children are vaccinated. Let me add, without going into too much detail, that</td>
<td></td>
<td></td>
<td>2020-01-09</td>
</tr>
<tr>
<td>CommCh</td>
<td>but public health depends on people’s use of the vaccines that are available to them, and that should be encouraged</td>
<td></td>
<td></td>
<td>2020-01-09</td>
</tr>
<tr>
<td>CommCh</td>
<td>also means supporting the development and deployment of vaccines: an important signal of our commitment will be when w</td>
<td></td>
<td></td>
<td>2020-01-13</td>
</tr>
</tbody>
</table>
Hansard

whole corpus

2B

word

L5

R5

Reset view
Choose a corpus or multiple corpora.

Choose tools and modify settings.

Enter search term(s), if KWIC, GraphColl, Whelk or Text are used.

Load search terms from a file

Modify the name of the report and its location

Run the tool
### 3.5 Statistical analysis

So far, descriptive statistical analysis was reported in the sections above. This includes the analysis of frequency and dispersion (sections 3.1 and 3.3), keywords (section 3.2) and collocations (section 3.4). More details about these procedures can be found in Brezina (2018).

The t-test ($t(28) = -0.82$, $p = 0.421$) showed that there is not a statistically significant difference between the corpora with regard to the linguistic variable *good*. This result is visualised below. Cohen's $d$ (-0.3, 95% CI [-1.05, 0.42]).
Advantages of #LancsBox

- More *time* for exploration.
- More *time* for multiple perspectives.
- More *time* for linguistic interpretation.
- More *time* for things that matter.
thank you much for god goodness you
AntConc:
A corpus analysis toolkit for concordancing and text analysis

Laurence Anthony
Center for English Language Education in Science and Engineering (CELESE)
Faculty of Science and Engineering, Waseda University
http://www.laurenceanthony.net/
anthony@waseda.jp
@antlabjp

Overview

- History of AntConc
  - release history
  - user statistics
- Overview of AntConc 4
  - design concept
  - tools and features
    - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, and Keyword, Wordcloud
- Live Demonstration
  - from corpora to meaningful KWIC results
  - from corpora to meaningful Keywords results
- Q&A

History of AntConc:
One of the 17 "AntLab" freeware language processing tools

Data Collection Tools
- AntCorGen
- AntFileConverter
- AntFileSplitter
- FireAnt
History of AntConc:
One of the 17 "AntLab" freeware language processing tools

Data Cleaning Tools
- EncodeAnt
- SarAnt
- SegmentAnt

Data Annotation Tools
- AntMover
- CLAWSAnt
- SegmentAnt
- TagAnt

History of AntConc:
One of the 17 "AntLab" freeware language processing tools

Data Analysis Tools
- AntConc
- AntPConc
- AntWordProfiler
- FireAnt
- ProtAnt
- VariAnt

History of AntConc:
Four major releases

- Version 1.0-1.1: Simple concordancer for Windows and Linux
  - tools: concordance
  - language: Perl + Tk
History of AntConc:
Four major releases

- Version 2.0-2.6: Corpus toolkit for Windows and Linux
  - tools: concordance, plot, cluster/n-grams, file, word, keyword
  - language: Perl + Tk

- Version 3.0-3.5: Corpus toolkit for Windows, Linux, and MacOS (native)
  - tools: concordance, plot, file, cluster/n-grams, collocates, word, keyword
  - language: Perl + Tkx

History of AntConc:
Four major releases

- Version 4.0-4.1: Advanced corpus toolkit for Windows, Linux, and MacOS (native)
  - tools: concordance, plot, file, cluster, n-grams, collocates, word, keyword, wordcloud
  - language: Python + Qt

History of AntConc:
User statistics

- corpus.by.edu
- AntConc
- WordSmith Tools
- Sketch Engine
- Sarah (with BNC)
- Monoconc Pro
- Xaira (with BNC XML or your own corpus data)
- WMMatrix
- Oxford Concordancing Program
- Longman Mini-concordancer
- Other

"Which computer programs do you use for analysing corpora?"
Overview of AntConc 4:
Design concept - strengths

- easy to find, set up, and use
  - e.g., installer or portable version (no installation required)
  - e.g., help guides, YouTube tutorial videos, discussion forums, ...
- safe and reliable
  - e.g., digitally signed and notarized by Windows and MacOS (no security warnings)
  - e.g., no viruses, community error reports, bug fixes, updates, ...
- managed by the designer
  - i.e., the project leader built the tool
- evolving
  - e.g., new tools, new features, new documentation, ...
- designed for a broad user base
  - e.g., multiplatform support (Win, Mac, Linux)
  - e.g., multilingual support (UTF-8 compliant)
  - e.g., supports LTR (e.g., English, Japanese) and RTL languages (e.g., Arabic)
Overview of AntConc 4:
Design concept - strengths

- comprehensive, scalable, fast, and transparent
- includes all standard corpus tools
  - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word (list), Keyword (list), Wordcloud
- includes access to an online corpus repository
  - e.g., AmE06, BE06, ... (BASE, BAWE, and more corpora coming soon)
- uses fully-indexed, single-file SQLite databases
  - corpora of hundreds, thousands, millions, and billions of words can be used
    - e.g., "the" searches for any corpus size are returned in milliseconds
    - e.g., results are "paged" to prevent endless scrolling
  - corpora can be loaded, saved, and distributed easily
  - corpora tables can be easily inspected, edited, and copied
- allows importing and exporting of results as plain text files
  - MTSV format (Anthony & Evert, 2019)
  - full data interoperability is possible

Overview of AntConc 4:
Design concept - weaknesses

- managed by the designer
  - i.e., the project leader built the tool
- designed for a broad user base
  - e.g., few niche tools and statistics are included
- a "black box" technology
  - e.g., the software is 'freeware' but not open-source'
  - (currently) inflexible
  - e.g., users cannot easily add new features or statistical methods

Overview of AntConc 4:
Tools and features - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword, Wordcloud

Overview of AntConc 4:
Tools and features - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword, Wordcloud
Overview of AntConc 4:
Tools and features - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword, Wordcloud

Overview of AntConc 4:
Tools and features - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword, Wordcloud

Overview of AntConc 4:
Tools and features - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword, Wordcloud

Overview of AntConc 4:
Tools and features - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword, Wordcloud
Live Demonstration:
From corpora to meaningful KWIC and keyword results

- AntConc "Corpus Manager" – online/offline repository + user corpus library

Live Demonstration:
From corpora to meaningful KWIC results

- KWIC results - value ordered (alphabetical to the right)

Live Demonstration:
From corpora to meaningful KWIC results

- KWIC results - value ordered (alphabetical to the left)
Live Demonstration:
From corpora to meaningful KWIC results

- KWIC results - frequency ordered (word patterns to the left)

```plaintext
Target Corpus
Name: Asnedk

<table>
<thead>
<tr>
<th>File</th>
<th>Cluster</th>
<th>N-Gram</th>
<th>Collocate</th>
<th>Word</th>
<th>Keyword</th>
<th>Target Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>File</td>
<td>left Context</td>
<td>Kit</td>
<td>Right Context</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. An else if present, let us see what余人 find about the study of details. In the late 1980s, Fud i held a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. An else, focus the chapters on a single approach to the study of leadership, but not to</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. An else, has identified a host of approaches to the study of leadership. What are the personal characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. An else, go in fact, there are now several distinct the</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. An else, is a revealing instance of views on life and thinking. How might the</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. An else, space, space, basis, and orthogonal basis in the study of linear models.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. An else, one of the most striking developments in the study of child language, and indeed developmental p</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. An else, there may be a statistical model in the study as a collaborator from the beginning. Most apps</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. An else, the authors of the study of the whole text. An idea of how they think of their</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. An else, refers to the study of the whole text. An idea of how they think of their</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. An else, refer to the study of the whole text. An idea of how they think of their</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. An else, is a revealing instance of views on life and thinking. How might the</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Search Query:
Words: Case
Regexp: Start
Order by Rank
```

Live Demonstration:
From corpora to meaningful keyword results

- Keyword results produced by Log-Likelihood

```plaintext
Target Corpus
Name: Asnedk

<table>
<thead>
<tr>
<th>Keyword Type</th>
<th>Type</th>
<th>Rank</th>
<th>Score</th>
<th>Collocate</th>
<th>Range,</th>
<th>Target,</th>
<th>Range,</th>
<th>Ref</th>
<th>Keywords</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. however</td>
<td>1</td>
<td>107</td>
<td>444</td>
<td>65, 254</td>
<td>39,102</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. thus</td>
<td>2</td>
<td>2</td>
<td>206</td>
<td>49, 138</td>
<td>37,563</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. study</td>
<td>3</td>
<td>120</td>
<td>312</td>
<td>51, 142</td>
<td>36,827</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. particular</td>
<td>4</td>
<td>75</td>
<td>197</td>
<td>47, 128</td>
<td>53,037</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. g</td>
<td>5</td>
<td>528</td>
<td>176</td>
<td>35, 65</td>
<td>32,299</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. studies</td>
<td>6</td>
<td>100</td>
<td>181</td>
<td>56, 60</td>
<td>30,695</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. such</td>
<td>7</td>
<td>219</td>
<td>1030</td>
<td>76, 346</td>
<td>30,095</td>
<td>0.004</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. examples</td>
<td>8</td>
<td>120</td>
<td>322</td>
<td>59, 132</td>
<td>29,727</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. e</td>
<td>9</td>
<td>182</td>
<td>243</td>
<td>63, 115</td>
<td>29,619</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. associated</td>
<td>10</td>
<td>60</td>
<td>106</td>
<td>39, 60</td>
<td>27,854</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Search Query:
Words: Case
Regexp: Start
Order by Rank
```

Live Demonstration:
From corpora to meaningful keyword results

- Keyword results produced by Text Dispersion

```plaintext
Target Corpus
Name: Asnedk

<table>
<thead>
<tr>
<th>Keyword Type</th>
<th>Type</th>
<th>Rank</th>
<th>Score</th>
<th>Collocate</th>
<th>Range,</th>
<th>Target,</th>
<th>Range,</th>
<th>Ref</th>
<th>Keywords</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. however</td>
<td>1</td>
<td>107</td>
<td>444</td>
<td>65, 254</td>
<td>39,102</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. thus</td>
<td>2</td>
<td>2</td>
<td>206</td>
<td>49, 138</td>
<td>37,563</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. study</td>
<td>3</td>
<td>120</td>
<td>312</td>
<td>51, 142</td>
<td>36,827</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. particular</td>
<td>4</td>
<td>75</td>
<td>197</td>
<td>47, 128</td>
<td>53,037</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. g</td>
<td>5</td>
<td>528</td>
<td>176</td>
<td>35, 65</td>
<td>32,299</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. studies</td>
<td>6</td>
<td>100</td>
<td>181</td>
<td>56, 60</td>
<td>30,695</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. such</td>
<td>7</td>
<td>219</td>
<td>1030</td>
<td>76, 346</td>
<td>30,095</td>
<td>0.004</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. examples</td>
<td>8</td>
<td>120</td>
<td>322</td>
<td>59, 132</td>
<td>29,727</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. e</td>
<td>9</td>
<td>182</td>
<td>243</td>
<td>63, 115</td>
<td>29,619</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. associated</td>
<td>10</td>
<td>60</td>
<td>106</td>
<td>39, 60</td>
<td>27,854</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Search Query:
Words: Case
Regexp: Start
Order by Rank
```

Summary
Summary

- **AntConc 4**...
  - is a freeware, multiplatform toolkit for doing data-driven language learning, introducing corpus methods, and carrying out corpus linguistics research.
  - runs on any computer running Windows, MacOS, and Linux.
  - includes **KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, and Keyword, Wordcloud** tools
  - uses a fully-indexed SQLite database for storing thousand-, million-, and billion-word corpora
  - offers access to an online repository of corpora

AntLab Tools
www.laurenceanthony.net/software/antconc
KonText: an online corpus query interface and its use in CLARIN

Michal Křen
Charles University, Prague

CLARIN Café
15 September 2022
## Introduction

### Desktop tools
- tools for people to download and use on their own computer
- typically also on their own data
- easy installation, (almost) no configuration, easy to run and use
- more limitations, especially on size and number of annotation layers

### Online tools
- tools available on the internet, registration usually required
- typically on pre-installed corpora
- professional installation and maintenance of the tools necessary
- powerful query language, scalability, ability to handle more rich annotation
## KonText

### Characteristics

- open-source web-based general-purpose corpus concordancer
- originally a fork of NoSketch Engine:
  - graphical user interface: eventually completely rewritten
  - backend: Manatee-open ⇒ KonText the same query language as (No)SkE
- main technical features:
  - modular design with plug-ins
  - structured application log & configuration files
  - continuous and open development (since 2014) at GitHub (with the needs of other centres in mind)
  - developers: Tomáš Machálek (lead) & Martin Zimandl
  - deployed/tested also at other CLARIN centres (Poland, Slovenia)
  - CLARIN Federated Content Search endpoint

- https://korpus.cz/kontext/
Most notable user features

before entering a query:
- only two query types: *simple* and *advanced*
  (*simple* mode is still highly configurable, *advanced* mode means full CQL)
- CQL editor with syntax highlighting and basic value validation
- tag builder widget for interactive selection of morphological categories
  (supports also Universal Dependencies)
- query history with an easy overview, filtering, and marking for later reuse

working with the query result:
- filtering and sorting (including random shuffling) of concordances
- statistical functions, collocation analysis
- frequency charts with confidence intervals (as an alternative to the tables)
- breadcrumb navigation as an editable processing chain
- each query (and respective result) has a public permanent ID
  ⇒ easily sharable results
Most notable user features II

▶ working with subcorpora:
  ▶ interactive creation of subcorpora by “zooming into” the selected parts of a corpus down to the document level
  ▶ contents of a subcorpus can be edited and used as a base for another subcorpus
  ▶ every subcorpus gets a public permanent ID

▶ supported corpus types and annotation layers:
  ▶ support for parallel corpora
  ▶ support for spoken corpora: visual representation of dialogues with indication of speaker turns and overlaps
  ▶ rendering of dependency syntax trees (syntactic queries still require CQL)
Search in the corpus

Create / edit a tag

Selected features:
Number = Plur & POS = NOUN

Part of speech:
- ADJ
- ADP
- ADV
- AUX
- CC
- CONJ
- DET
- INTJ
- NOUN
- NUM
- PART
- PRON
- PROPN
- PUNCT
- SCONJ
- SYM
- VERB
- X

Features:
- Plur
- Sing
- Abbr (0)
- Case (0)
- Definite (0)
- Degree (0)
- Foreign (0)
- Gender (2)
- Mood (1)
- Number (2)
- NumType (0)
- Person (2)
- Polarity (0)
- Poss (0)
- PronType (2)
- Reflex (0)
- Tense (1)
- Typo (0)
- VerbForm (1)

Insert
Undo
Reset
Search in the corpus

InterCorp v13ud - English

Advanced query | Insert tag | Insert within | Keyboard | Recent queries

[upos="ADJ"] []{0,3} [feats="Number=Plur" & upos="NOUN" & deprel="nsubj"]

TIP You can click a tag value while holding CTRL to edit the tag using an interactive tool (next tip)

Specify parameters

Default attribute: word

Aligned corpora

Specify context

Restrict search

Search
there is every reason to suppose that Phrygian or later Phoenician merchants reached the shores of Albion.

results from this, it is even possible that as local emissions fall, global emissions will rise.

battle in which the British were successful. The remaining Roman troops, fearing massacre, fled with young couple today ... complaints by neighbors led them to disperse.

to a war... Police in Dallas arrested a few hours without her nerves betraying her, until her grand...

released at once to return to Scotland. I am sure your families are missing you greatly at this national average annual transboundary fluxes or national average annual...

... to any party specifying such a previous year, its national average annual transboundary fluxes or national average annual e...

Listen. Thank you. Nonsense. What are national average annual transboundary fluxes or national average annual e...

to a halt as a result. There are two other areas where we must make decisio...

the heels of monstrous death. But in a more elusive way the streets and houses of that vicinity b...

the London world, particularly in the novels of the twentieth century, where feelings of hopelessness and despair...

of the City. In the early decades of the twentieth century new office blocks became a familiar presence.

been the case for thousands of years and, in earlier times, some activities — extensive land-use practi...

I could hardly forget the lost center when all my open wounds cry out for the unity that grou...
# The remaining Roman troops, fearing massacre, fled within the walls and closed the gates.
<table>
<thead>
<tr>
<th>Corpus: InterCorp v13ud - English</th>
<th>Query: ADJ, Number=Plur, NOUN, nsubj (207,246 hits)</th>
<th>InterCorp v13ud - German</th>
</tr>
</thead>
<tbody>
<tr>
<td>□ Acquis</td>
<td>Appropriate <strong>internal and external financial controls</strong> are to be carried out in accordance with internationally accepted audit standards by the competent financial control authority, which must be independent for the purposes of performing this function. &lt;/S&gt;</td>
<td>□ Acquis</td>
</tr>
<tr>
<td>□ Core</td>
<td>But though the phalanx stayed more or less constant, the <strong>great beasts</strong> of which it was composed charged steadily at upwards of twenty miles an hour, appearing suddenly from thin air at one end of the plain, and disappearing equally abruptly at the other end. &lt;/S&gt;</td>
<td>□ Core</td>
</tr>
<tr>
<td>□ Europarl</td>
<td>I voted for the report because I believe that these <strong>technological advances</strong> need to be adequately regulated in order that they do not cause systemic risk to the overall functioning of the markets. &lt;/S&gt;</td>
<td>□ Europarl</td>
</tr>
<tr>
<td>□ Core</td>
<td>We are living in a time when flowers are trying to live on flowers, instead of growing on good rain and <strong>black loam</strong>. &lt;/S&gt;</td>
<td>□ Core</td>
</tr>
<tr>
<td>□ Core</td>
<td>Its <strong>myriad facets</strong> sparkled as only pricey crystal can and splashed miniature rainbows onto the walls. &lt;/S&gt;</td>
<td>□ Core</td>
</tr>
<tr>
<td>□ Core</td>
<td><strong>Muslim-majority countries</strong> spend, on average, less than 0.5% of their GDP on research and development, compared with five times that in the advanced economies. &lt;/S&gt;</td>
<td>□ Core</td>
</tr>
<tr>
<td>□ Syndicate</td>
<td>Länder mit muslimischen Bevölkerungsmehrheiten geben durchschnittlich weniger als 0.5% ihres BIP für Forschung und Entwicklung aus; &lt;/S&gt;</td>
<td>□ Syndicate</td>
</tr>
</tbody>
</table>
KonText user interface: depiction of utterances and overlaps in a dialogue.
Thank you for your attention!