

How to Search for Linguistic Patterns across Corpora

CL2023: “What can you do with the CLARIN infrastructure?”

Tutorial Handout

Iulianna van der Lek (CLARIN ERIC), Martin Wynne (University of Oxford)

The aim of this tutorial is to show how to use the CLARIN [Content Search](#) (also known as Federated Content Search) service to search for specific patterns across collections of data. This service can be used in the first phase of a (research) project, for example, to locate interesting language resources across distributed collections of digital texts.

Workload

- **Practical exercise:** 20 min (estimate)

LEARNING OUTCOMES

By the end of this lesson, students will learn to:

- Define Federated Content Search (FCS)
- Explain the difference between federated content search and metadata search
- Use FCS to locate specific patterns across data collections
- Process the search results with more advanced tools


TOPICS

- What is Content Search and why would you use it?
- How to access, search, view and download the search results
- How to process and analyse the search results with more advanced tools

What is Content Search and Why Use it?

The CLARIN-FCS, a **common federated content search infrastructure service**, can be used to connect to the local data collections available in the CLARIN specialised centres, perform queries within the resource content and analyse the search results.

This service is **useful in the first stage of a (research) project when** you need to locate resources and datasets available across several national repositories that might help you answer your research questions.

 The federated content search differs from the **metadata search** you performed in the VLO, where all metadata is harvested and then centrally indexed.

Access

FCS Aggregator is available at <https://contentsearch.clarin.eu/>. No login is required.

Basic Search

To perform a basic search within the available text collections, enter a search term and click the magnifying glass button or press **Enter**.

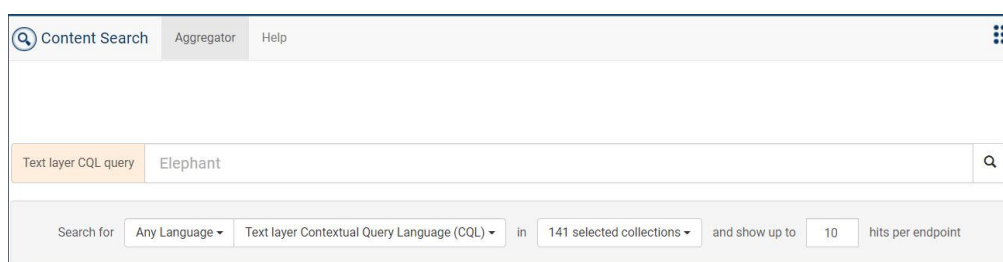



Figure 1. The search interface of Content Search

The default search options will apply: all the available FCS collections in any language will be searched and up to 10 hits per collection will be displayed.

 By default, 10 records per collection are returned. You can change this number by using the right-most control of the control bar. More records can be retrieved at any time, in the focused view, when available.

Search options

The following search options are available:

Option 1: Specify the language of the resource

1. Click on **Any Language** to access the dialogue box with multiple language options.
2. Select a language from the list and try out one of the filtering options:
 - Use the specified language of the collection to search within the collection that is explicitly designated for the chosen language.
 - Search across all collections using a language guesser to determine the language of each resource; only the results that align with the selected language will be displayed.
 - Search within collections that contain resources in the selected language, then apply the language guesser to further refine the results.



Figure 2. Select the language in FCS

Option 2: Specify the collections to search in

By default, all the corpora collection sources are selected. If you want to restrict your search to a specific set of resources, click on the middle button in the control bar, initially displaying **All available collections** and a dialogue window will show up. This dialogue provides the option to select and deselect all collections, or particular collections.

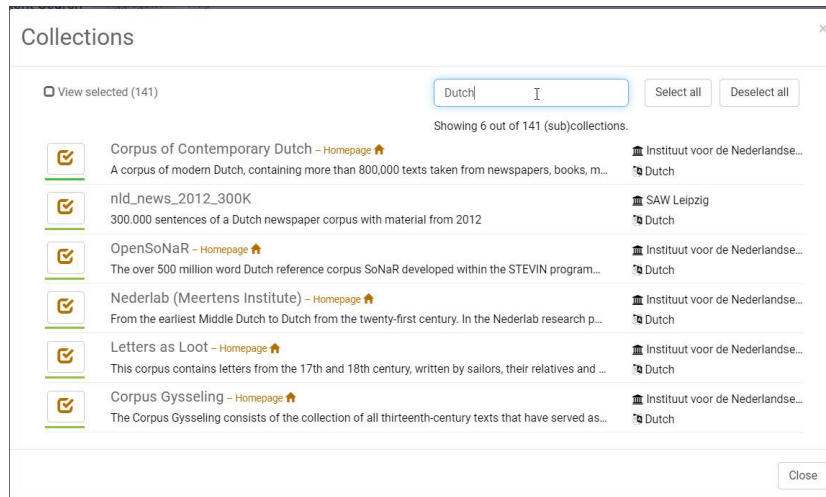


Figure 3. List of collections in FCS

Some collections also have sub-collections; in this case, there will be a link to expand and explore, select and deselect the sub-collections.

Viewing the search results

As soon as the search function is invoked, the query is sent to all the selected corpora and the search results from resources are displayed dynamically as soon as they come in. Therefore, you can start inspecting the search results (the records) before all the corpora have returned a response. A progress bar at the top of the page provides information about the status of the search.

The search results are initially displayed in a textual view.

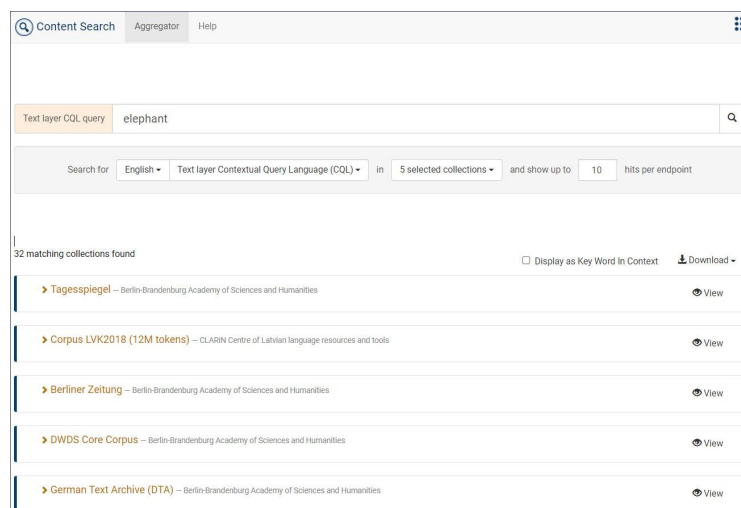


Figure 4. View the FCS search results

The textual view can be toggled to a concordance view by clicking on **Display as Key Word in Context**:

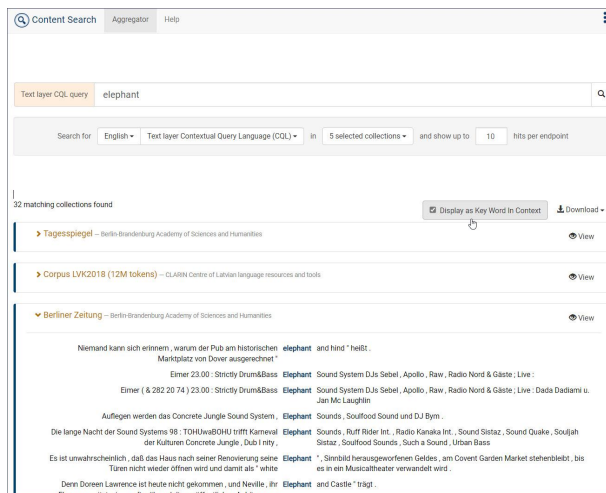


Figure 5. Display the search results in FCS as a keyword in the context

Download search results

You can download the entire set of search results onto your computer in one of the following formats: CSV, ODS, Excel, TCF, or a plain text file.

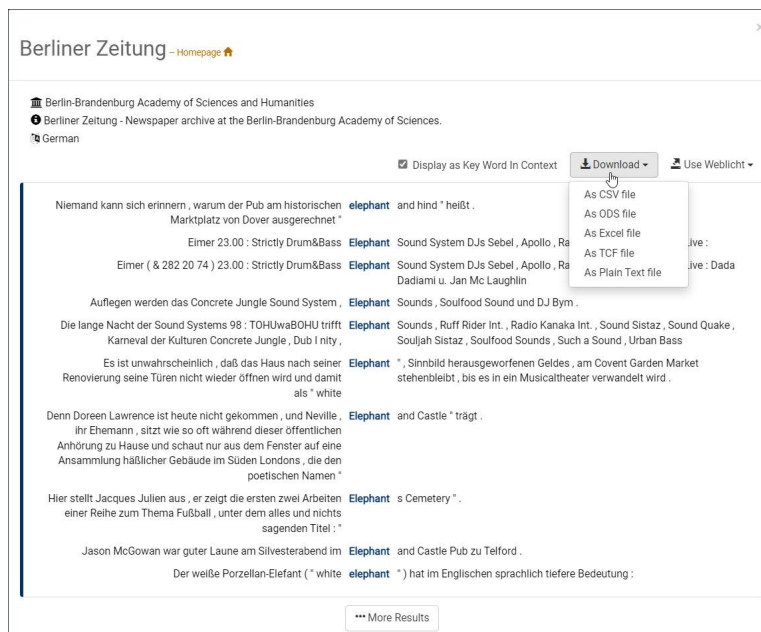


Figure 6. How to download the search results in FCS

Process the search results

For a more in-depth analysis, you can send the search results directly to [WebLicht](#). If the service does not work, you can download the search results and upload the text to WebLicht manually to perform a more sophisticated analysis.

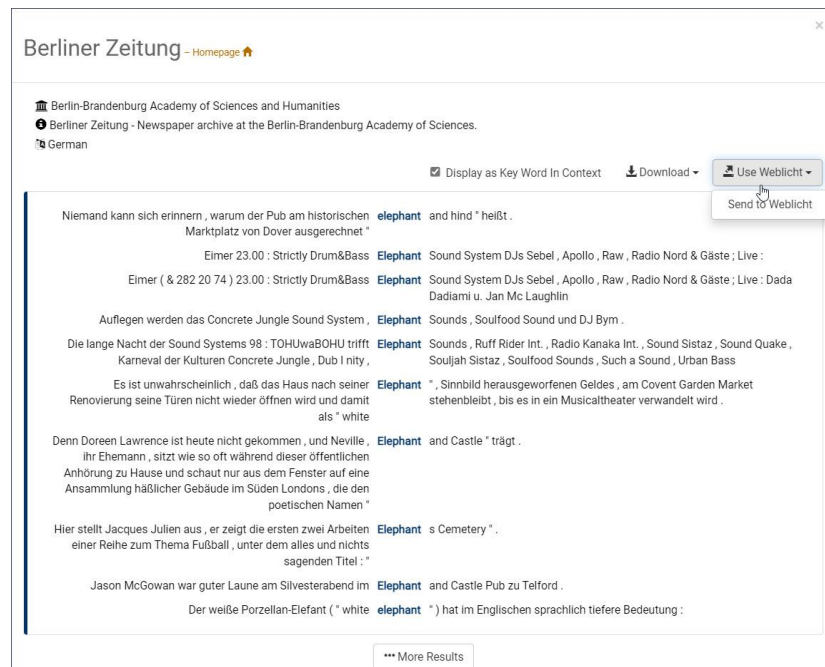


Figure 7. How to send the search results in FCS to Weblicht

Alternatively, you can upload the downloaded plain text file to the [Language Resource Switchboard](#) to find a matching tool for another type of processing.

The examples below show processing the FCS search results with [UDPipe](#), a pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files. UDPipe is a language-agnostic tool and can be trained given annotated data in [CoNLL-U format](#).

Step 1: Upload the downloaded FCS results in plain text format to the Language Resource Switchboard.

Step 2: Switchboard will automatically suggest a list of tools that you can process the text. Select the **Dependency Parsing tool**, **UDPipe** and click **Open**.

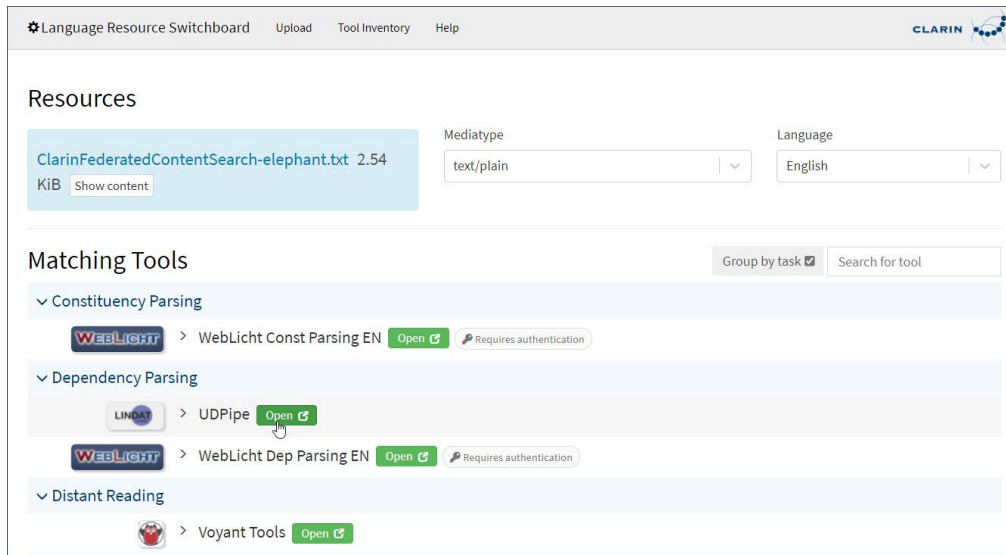


Figure 8. How to access UDPipe from Switchboard

Step 3: You can view and download the output file for further analysis.

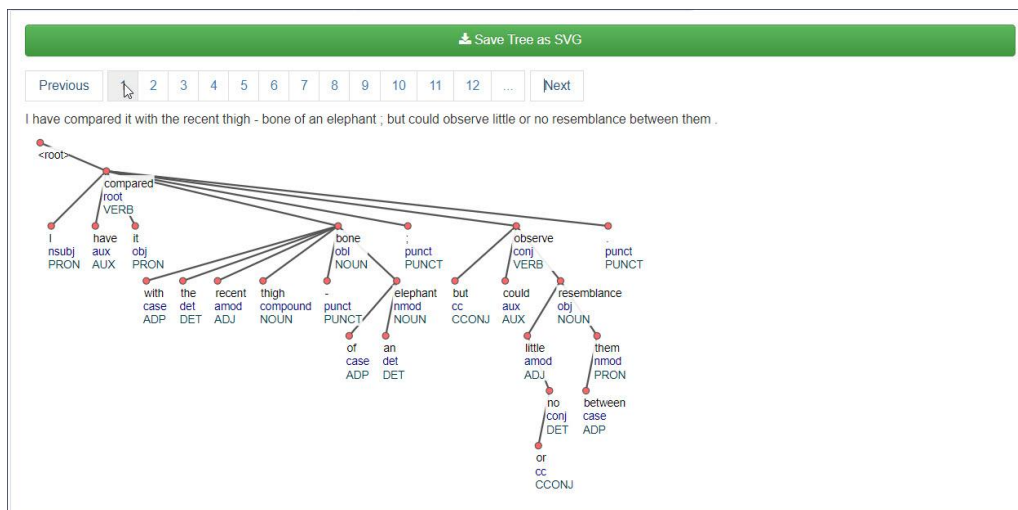


Figure 9. View the tree in UDPipe

Choosing a corpus

Often users don't want to do things with the aggregated results, but are more interested in using the Content Search service in order to find the right corpus or corpora to investigate further. From any of the results produced by a Content Search query, you can click on View and then the 'Homepage' link next to the corpus name at the top to be redirected to the interface for that online corpus.

Es markiert das Ankommen der maghrebinischen **Migranten** in der französischen (Pop-)Kulturturnation .

» Für einige Gemeinden der Mixteca Poblana organisierten sich in New York Unterstützungskomitees , die sich z. B. die Verlegung von Trinkwasserleitungen in ihren Herkunftsgemeinden oder die Restauration von Kirchen und Dorfplätzen zum Ziel setzten und hierfür bei den in New York arbeitenden **Migranten** Spendensammlungen durchführten .

Neben den Sportverbänden der **Migranten** in New York unterstützt die Botschaft aktiv die Entwicklung von Guadelupana-Gruppen , die den Kult um die Jungfrau von Guadelupe (der wichtigsten mexikanischen Nationalheiligen) in New York organisieren sollen .

Ganze Straßenzüge (etwa der nördliche Teil der Amsterdam-Street oder neighbourhoods in Queens) zeugen von dieser inzwischen sehr stabilen Infrastruktur , auf die transnationale **Migranten** bauen können und die gleichzeitig durch sie reproduziert wird .

Das eigentliche Problem sind jedenfalls nicht die künftigen **Migranten** , sondern pendelnde Schwarzarbeiter - und die sind , auch ohne EU-Mitgliedschaft ihrer Heimatländer , längst in Westeuropa angekommen .

⋮ More Results

Close

Figure 10: Go to the corpus homepage

Using the Content Search in this way helps users to find out which corpora might have content useful for addressing their research questions, without having to visit lots of different sites and repeatedly submit the similar queries.

When you leave to visit the new site, it might involve registering or logging in for the site, depending on how it is set up. Some CLARIN interfaces require authentication, and might require you to have credentials from a University or a CLARIN login. If you don't have a University login that works for logging in, you can request a CLARIN user id from <https://idm.clarin.eu/home/> . It is important to fill in the section 'Motivation', and you can say that you are following exercises from the CL2023 workshop.

References

- [FCS Aggregator – Content Search \(clarin.eu\)](#)
- [CoNLL-U Format \(universaldependencies.org\)](#)
- [Language Resource Switchboard \(clarin.eu\)](#)
- [UDPipe \(cuni.cz\)](#)
- [WebLicht](#)
- [CLARIN Identity Provider](#)