

# CLARIN and Corpus Linguistics



**Martin Wynne**

[martin.wynne@ling-phil.ox.ac.uk](mailto:martin.wynne@ling-phil.ox.ac.uk)

Faculty of Linguistics, Philology and  
Phonetics, University of Oxford

**National Coordinator, CLARIN-UK**

***Teaching and Language Corpora  
2022***

University of Limerick

Wednesday 13<sup>th</sup> July 2022

<https://www.clarin.eu/event/2022/teaching-clarin-workshop-talc-2022>

# Back to Limerick...

## Oxford Text Archive Corpus Linguistics Resources

*Eurocall 2003 Workshop  
University of Limerick*

<http://www.ota.ox.ac.uk/>

Wednesday 1<sup>st</sup> September 2003

### Course Info

Workshop organisers:

Ylva Berglund, [Oxford Text Archive](#)

[ylva.berglund@oucs.ox.ac.uk](mailto:ylva.berglund@oucs.ox.ac.uk)

Pernilla Danielsson, [Centre for Corpus Linguistics](#), Birmingham University

[pemilla@ccl.bham.ac.uk](mailto:pemilla@ccl.bham.ac.uk)

Martin Wynne, [Oxford Text Archive](#)

[martin.wynne@oucs.ox.ac.uk](mailto:martin.wynne@oucs.ox.ac.uk)

### Programme

9:00 Arrival, registration, introduction

9:15 **Warm-up exercises**

10:00 **Talk: Corpus Resources for Language Learning**

10:30 Coffee break

11:00 **Hands-on session**

12:00 **Feedback from exercises and discussion**

12:30 Lunch

14:00 **Parallel Corpora**

15:30 Tea break

16:00 **Parallel Corpora**

17:00 End of seminar

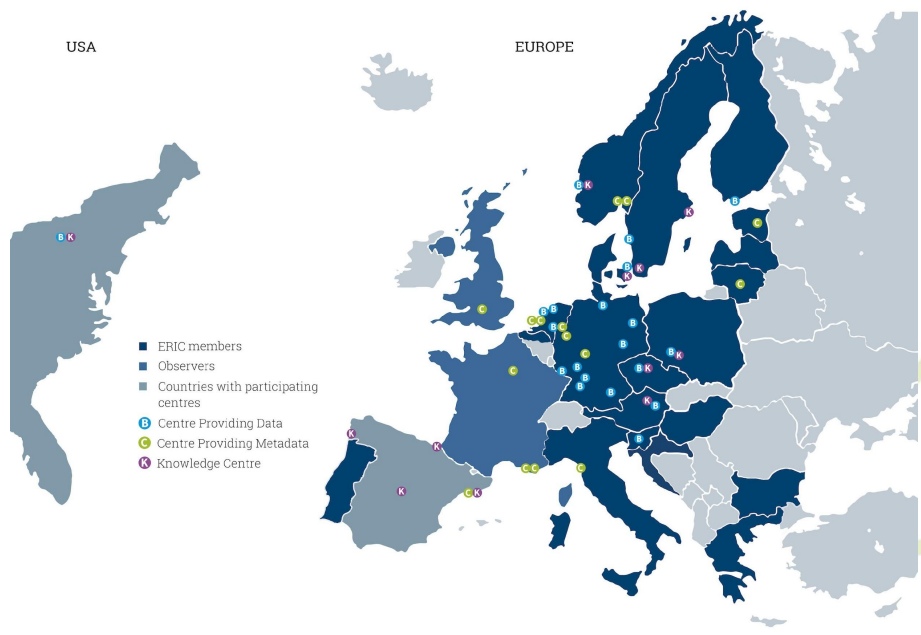
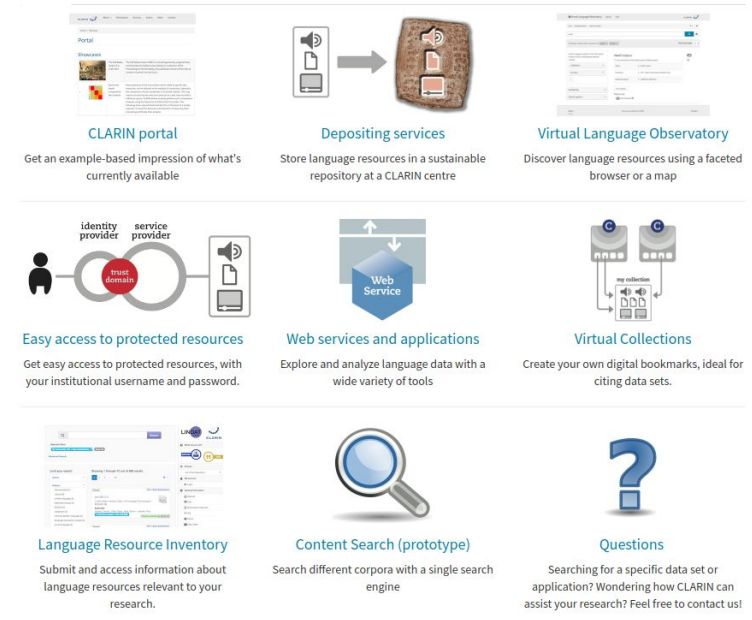
### Warm-up exercises

#### Missing keyword

What's the missing word in the concordance files on the handout sheet? Try not to shout out the answer while other people are still thinking about it!

How did you work out what the word was? Were there some lines which told you that there could only be one answer? Why? What different parts of speech does the missing word have in each line?

- **CLARIN in a nutshell**
- **CLARIN** is the Common Language Resources and Technology Infrastructure;
- it provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
- to **digital language data** (in written, spoken, video or multimodal form),
- and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
- through a **single sign-on** online environment.



### What's it all about?

- Discovery
- Support
- Sustainability
- Interoperability
- Cross-border collaboration

**CLARIN**  
**RESOURCE FAMILIES**



**Knowledge centres**

# What can CLARIN do for corpus linguists?

## For researchers

- Access to resources
- Networking opportunities
- Funding opportunities including for short visits, workshops, demonstrator and pilot projects,

## For teachers and learners

- [Teaching with CLARIN](#)
- [DH Course Registry](#)
- Access to resources
- Funding for creating teaching resources from CLARIN resources
- Funding for workshops
- Support for sharing and sustaining your tools and datasets
- Support for creating courses and learning objects from CLARIN resources

## For corpus creators

- Research Data Management for language corpora, including:
  - Repository services where data can be deposited
  - Access control and security
  - Long-term preservation
  - Discoverability (via the Virtual Language Observatory)
  - Interoperability (work in progress!)
  - Support for citation
  - Licence selectors
  - Support legal and ethical questions (GDPR, copyright, FAIR)
- Interoperability:
  - Discoverability: [Virtual Language Observatory](#)
  - Cross searching: [Federated Content Search](#)
  - Connecting data and tools: [Language Resources Switchboard](#)
- Connection to European networks and emerging best practice for data storage and management (EUDAT), accessing computing power (SSHOC), open access (FAIR), skills training (UPSKILLS), etc.

# 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](#)

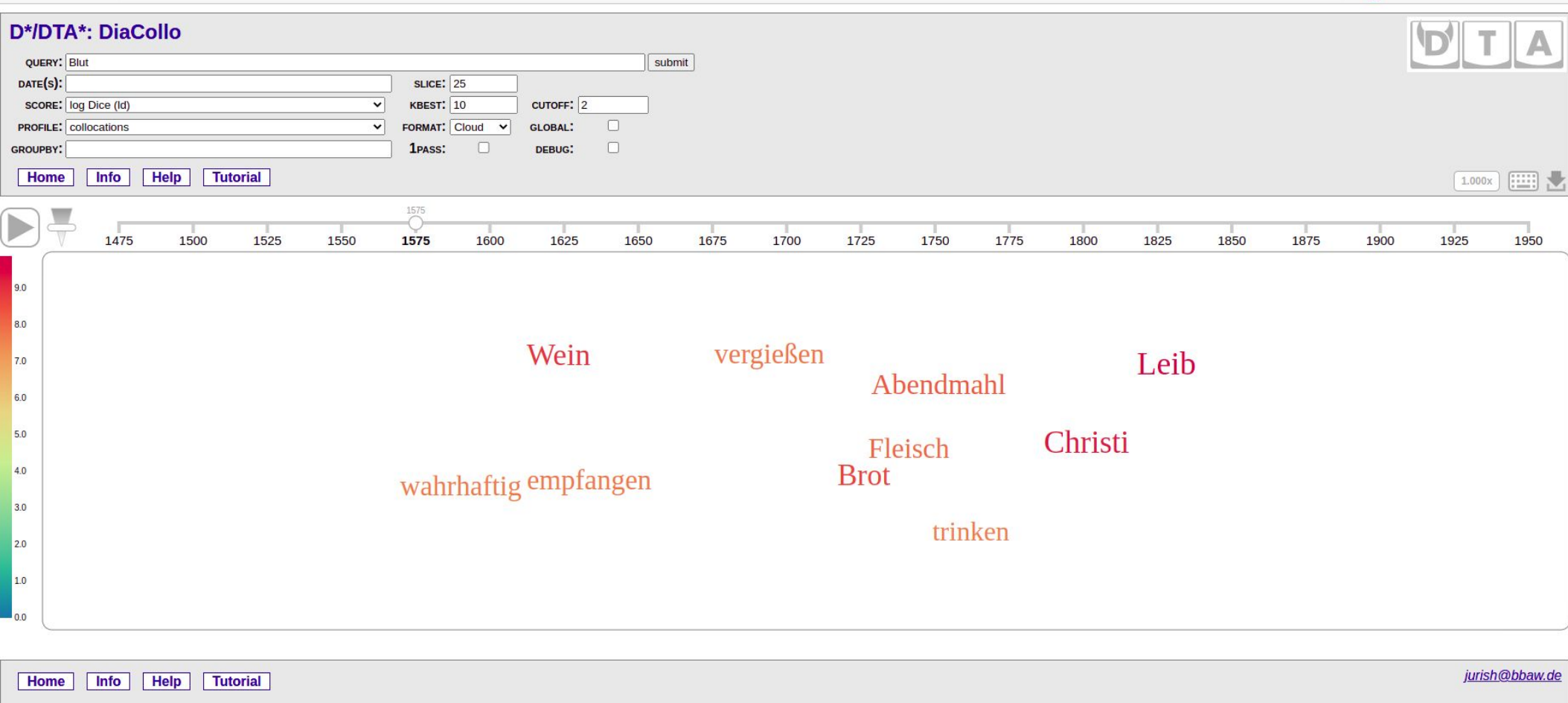


Federated content search: <https://contentsearch.clarin.eu/>

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





# D\*/DTA\* Search

Hits 1 - 10 of 498

[-HTML](#) [-Hist](#) | [Home](#) [Query Lizard](#) | [Previous](#) [Next](#) | [Help](#)    [CabErr](#)

- 1: [\[dtae:magirus\\_widerlegung\\_1580:8\]](#) ... Altars / daß daselbst warhafftig der Leib vnnd **Blut**<sub>[1]</sub> im Brot vnd **Wein**<sub>[2]</sub> / werde mündtlich geessen...
- 2: [\[dtae:magirus\\_widerlegung\\_1580:18\]](#) ... Brot vnd **Wein**<sub>[2]</sub> / seinen warhafftigen Leib vnd **Blut**<sub>[1]</sub> / zuessen vñ zutrincken darreiche vnd gebe:...
- 3: [\[dtae:magirus\\_widerlegung\\_1580:18\]](#) ... die vngläubige vnd vnwürdige / den Leib vñ **Blut**<sub>[1]</sub> Christi mit Brot vnd **Wein**<sub>[2]</sub> / im heiligen...
- 4: [\[dtae:magirus\\_widerlegung\\_1580:19\]](#) ... **Wein**<sub>[2]</sub> / aber gar nicht der Leib vnd **Blut**<sub>[1]</sub> Christi / gegenwertig dargereicht vñ empfangen werde /...
- 5: [\[dtae:magirus\\_widerlegung\\_1580:31\]](#) ... wir vns (ob der ware Leib vnd **Blut**<sub>[1]</sub> Christi leiblich im Brot vnd **Wein**<sub>[2]</sub> sey)...
- 6: [\[dtae:magirus\\_widerlegung\\_1580:31\]](#) ... anders gewölt / dann daß Christi Leib vnd **Blut**<sub>[1]</sub> war hafftig im Nachtmal mit Brot vnd **Wein**<sub>[2]</sub>...
- 7: [\[dtae:magirus\\_widerlegung\\_1580:35\]](#) ... / von einer wesentlichen Gegenwertigkeit des Leibs vnd **Bluts**<sub>[1]</sub> Christi / vnder gestalt des Brots vnd **Weins**<sub>[2]</sub>...
- 8: [\[dtae:magirus\\_widerlegung\\_1580:36\]](#) ... Brots vnd **Weins**<sub>[2]</sub> / in den Leib vnnd **Blut**<sub>[1]</sub> Christi (also daß kein Brot vnd Wein...
- 9: [\[dtae:magirus\\_widerlegung\\_1580:36\]](#) ... Brots vnd Weins / in den Leib vnnd **Blut**<sub>[1]</sub> Christi (also daß kein Brot vnd **Wein**<sub>[2]</sub>...
- 10: [\[dtae:magirus\\_widerlegung\\_1580:37\]](#) ... / vnd mit dem sichtbarn **Wein**<sub>[2]</sub> / das **Blut**<sub>[1]</sub> Christi dargereicht werde.

[-HTML](#) [-Hist](#) | [Home](#) [Query Lizard](#) | [Previous](#) [Next](#) | [Help](#)    [CabErr](#)

[jurish@bbaw.de](mailto:jurish@bbaw.de)

D\* OpenSearch API version 0.60

[Imprint](#) · [Privacy](#)

0.693405 sec

Corpus: [Deutsches Textarchiv](#) (DTA-Kernkorpus + DTA-Erweiterungen)

Hilfe

Texte ▾

Projekt ▾

Dokumentation ▾

Impressum

 in den Titeldaten  im Korpus  in der Dokumentation

Magirus, Johannes: Gründliche Widerlegung des unwahrhaftigen zwinglischen Buches, welches Magister Ambrosius Wolf unter dem Titel (Historie der Augsburger Konfession) in öffentlichem Druck ausgehen lassen. Tübingen, 1580.

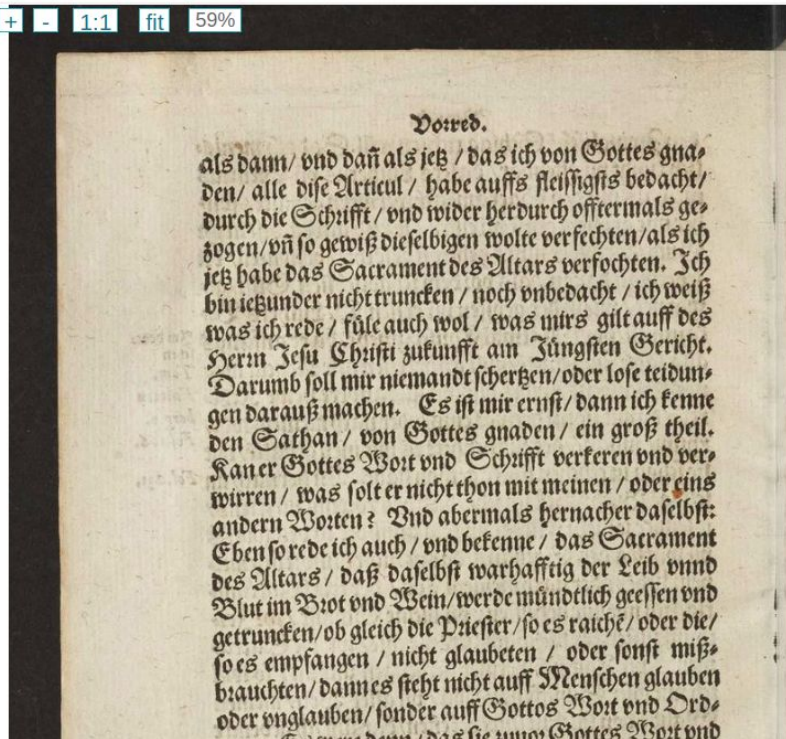
Bild: 0008 ▾

&lt;&lt; vorherige Seite

nächste Seite &gt;&gt;



+ - 1:1 fit 59%



als dann / vnd daß als jetz / das ich von Gottes gnaden / alle dise Articul / habe auffß fleissigsts bedacht / durch die Schrifft / vnd wider herdurch offtermals gezogen / vñ so gewiß dieselbigen wolte verfechten / als ich jetz habe das Sacrament des Altars verfochten. Ich bin ietzunder nicht truncken / noch vnbedacht / ich weiß was ich rede / füle auch wol / was mirs gilt auff des Herrn Jesu Christi zukunfft am Jüngsten Gericht. Darumb soll mir niemandt scherzen / oder lose teidungen darauß machen. Es ist mir ernst / dann ich kenne den Sathan / von Gottes gnaden / ein groß theil. Kaner Gottes Wort vnd Schrifft verkeren vnd verwirren / was solt er nicht thon mit meinen / oder eins andern Worten? Vnd abermals hernacher daselbst: Eben so rede ich auch / vnd bekenne / das Sacrament des Altars / daß daselbst warhafftig der Leib vñnd Blut im Brot vnd Wein / werde mündtlich geessen vnd getruncken / ob gleich die Priester / so es raichē / oder die / so es empfangen / nicht glaubeten / oder sonst mißbrauchten / dann es steht nicht auff Menschen glauben oder vnglauben / sonder auff Gotted Wort vnd Ordnung. Es were denn / das sie zuuor Gottes Wort vnd Ordnung endern / vnd anders deutten / wie die jetzigen Sacraments Feinde thun / wölche freilich eitel Wein vnd Brot haben / dann sie haben auch die Wort / vnd eingesetzte Ordnung Gottes nicht / sonder dieselbige nach jhrem eignen duncken verkert vñnd verendert.



# 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

# Contacts

National contacts:

See <https://www.clarin.eu/content/participating-consortia>

Contact the organisers here:

[francesca.frontini@ilc.cnr.it](mailto:francesca.frontini@ilc.cnr.it)

[martin.wynne@ling-phil.ox.ac.uk](mailto:martin.wynne@ling-phil.ox.ac.uk)

[i.vanderlek@uu.nl](mailto:i.vanderlek@uu.nl)

CLARIN K-Centres:

See <https://www.clarin.eu/content/knowledge-centres>

General contact:

[clarin@clarin.eu](mailto:clarin@clarin.eu)