Online and Desktop Tools for Querying a Language Corpus

CLARIN ERIC

15 September 2022

CLARIN

CLARIN-UK

CLARIN
Resource Families
Project Funding

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

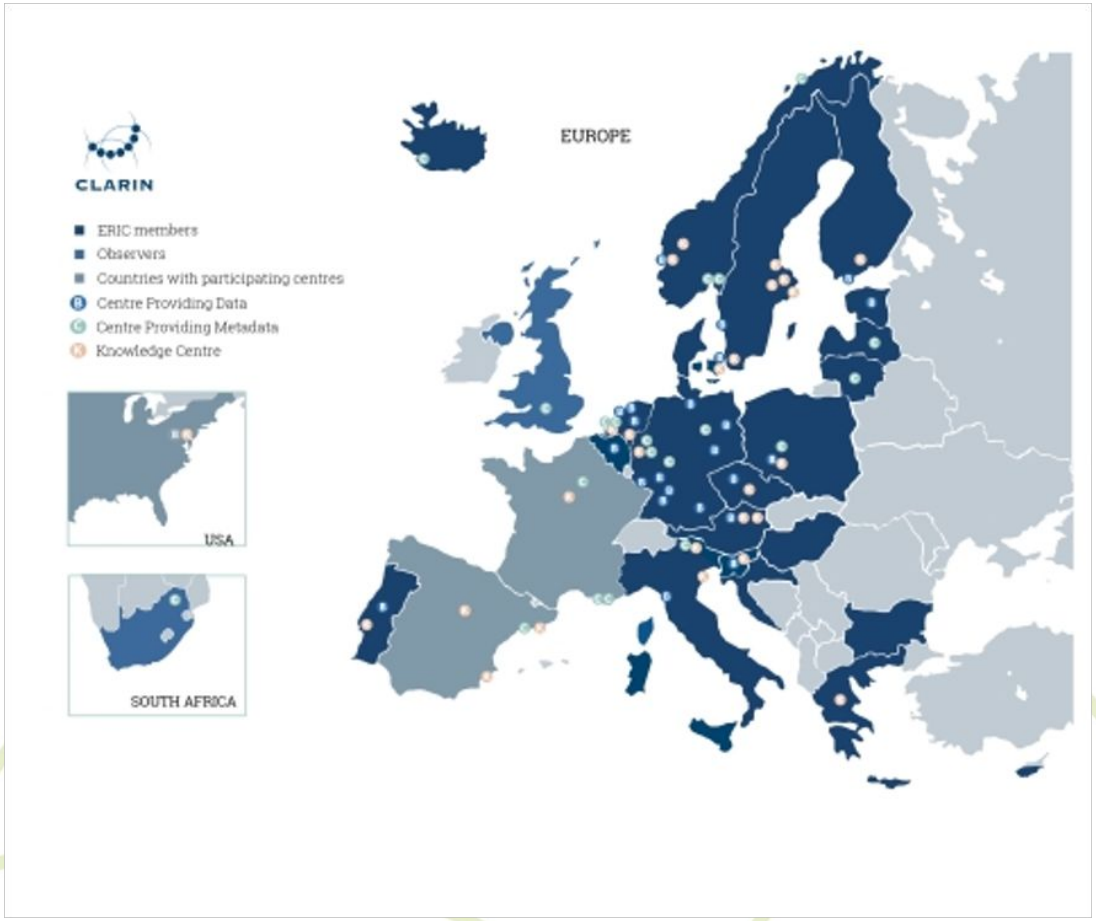
https://www.clarin.eu/content/clarin-in-a-nutshell

CLARIN …

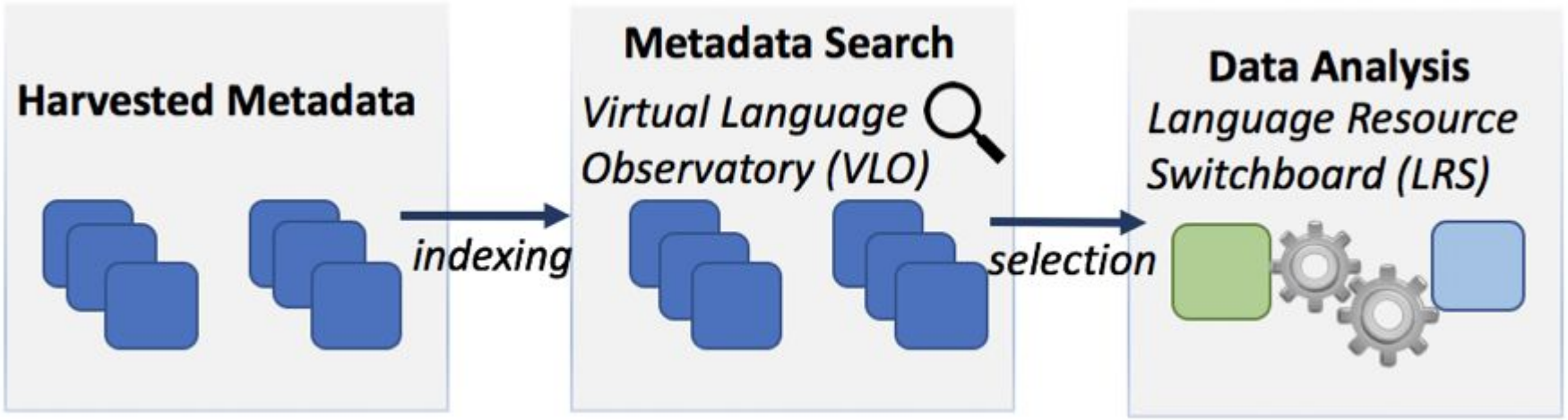
- is the *Common Language Resources and Technology Infrastructure*
- has the **ESFRI** ERIC status since 2012, Landmark since 2016
- provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
  - to **digital language data** (in written, spoken or multimodal 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**

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



Findable Accessible Interoperable Reusable

clarin.eu/fair

vlo.clarin.eu

switchboard.clarin.eu

The Knowledge Infrastructure

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:

https://www.clarin.eu/event/2022/clarin-annual-conference-2022

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

https://www.clarin.eu/event/2022/clarin-cafe-online-and-desktop-tools-querying-language-corpus

UK Research and Innovation

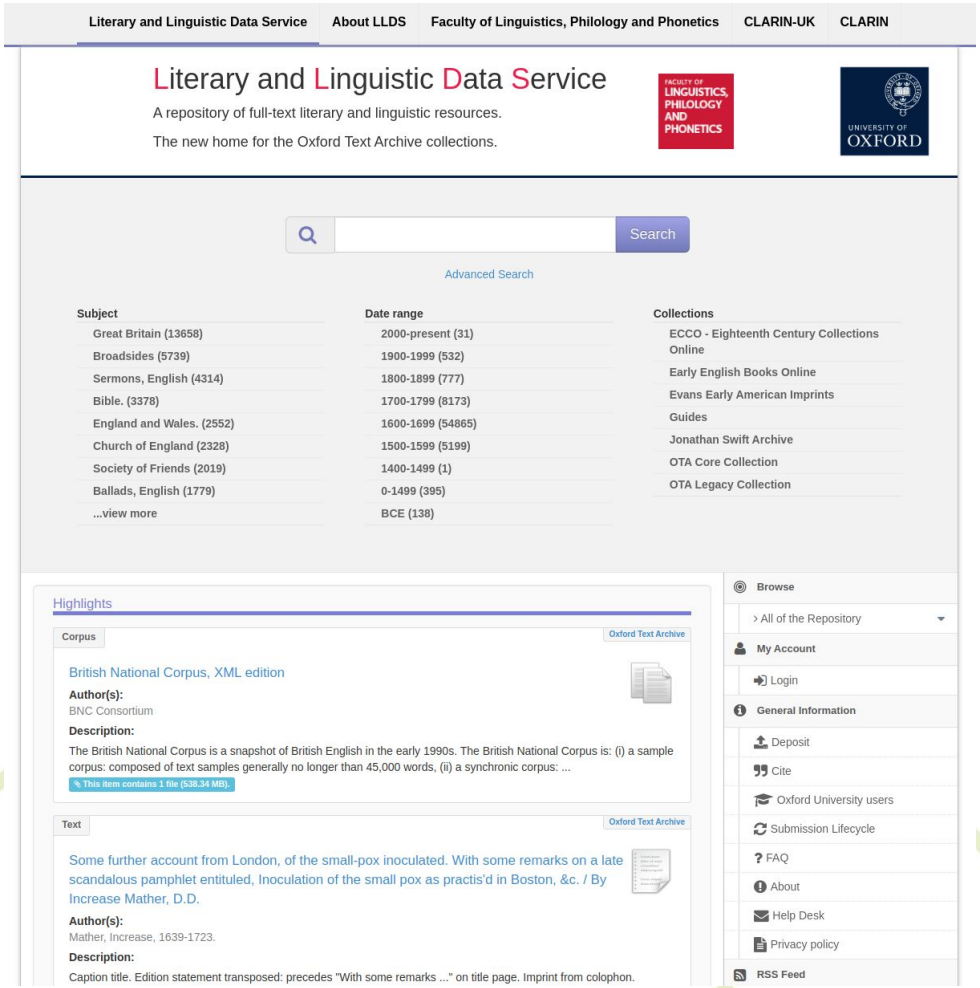
Home > Blog > Arts and humanities infrastructure enabling knowledge with impact

Arts and humanities infrastructure enabling knowledge with impact

Tao Chang
Head of Infrastructure, Arts and Humanities Research Council

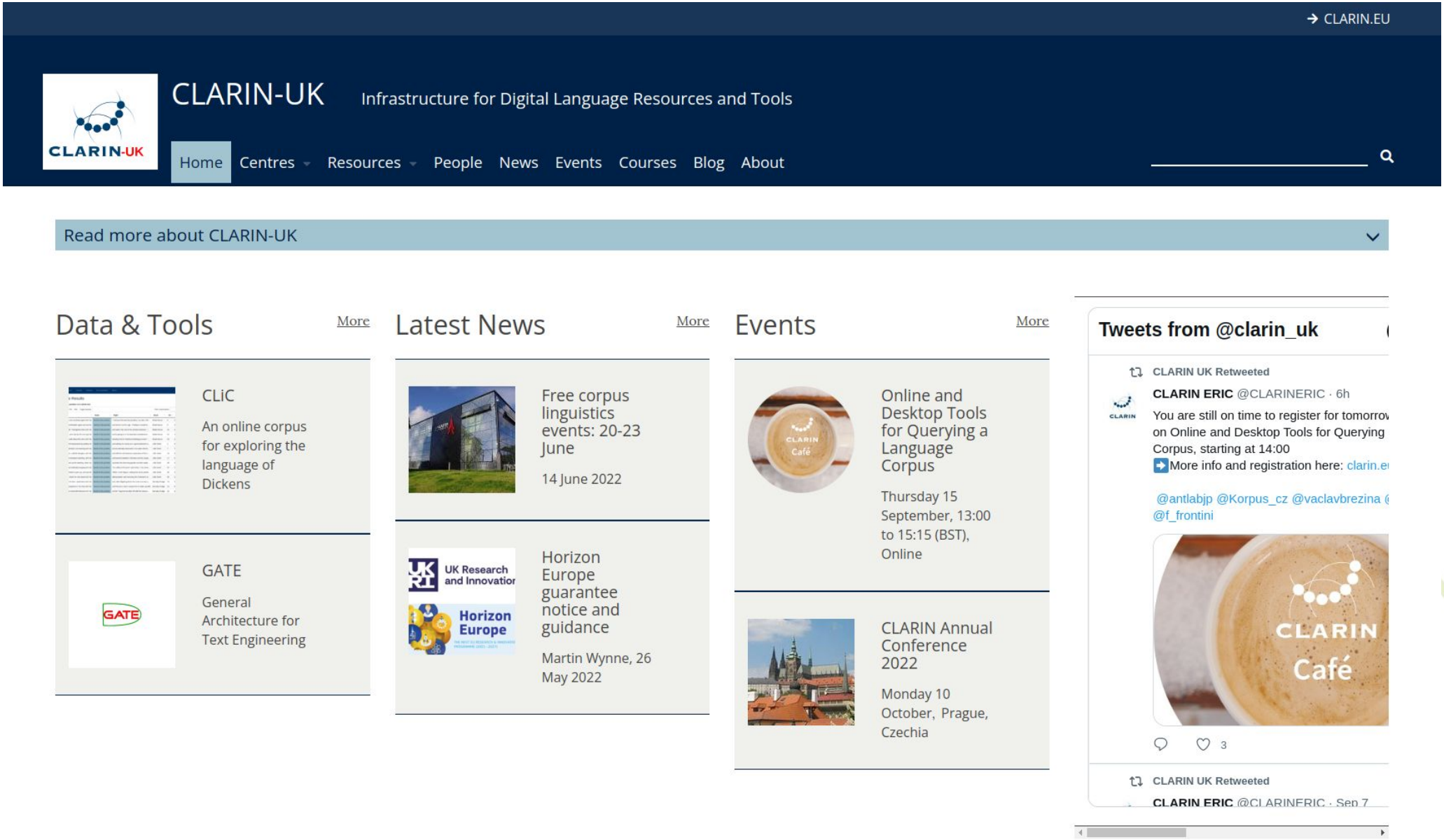
1 September 2021

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/

https://www.ukri.org/blog/arts-and-humanities-infrastructure-enabling-knowledge-with-impact/



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

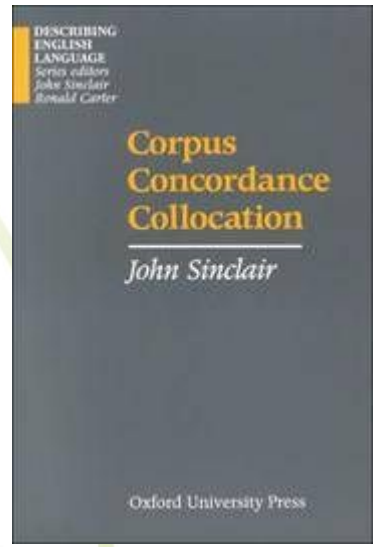
Coming soon:

- Online Corpus Tools
- Desktop Corpus Tools

https://www.clarin.eu/resource-families

Online and desktop tools for corpus analysis

"Corpus, concordance, collocation"



CONCORDANCE

LatinISE corpus

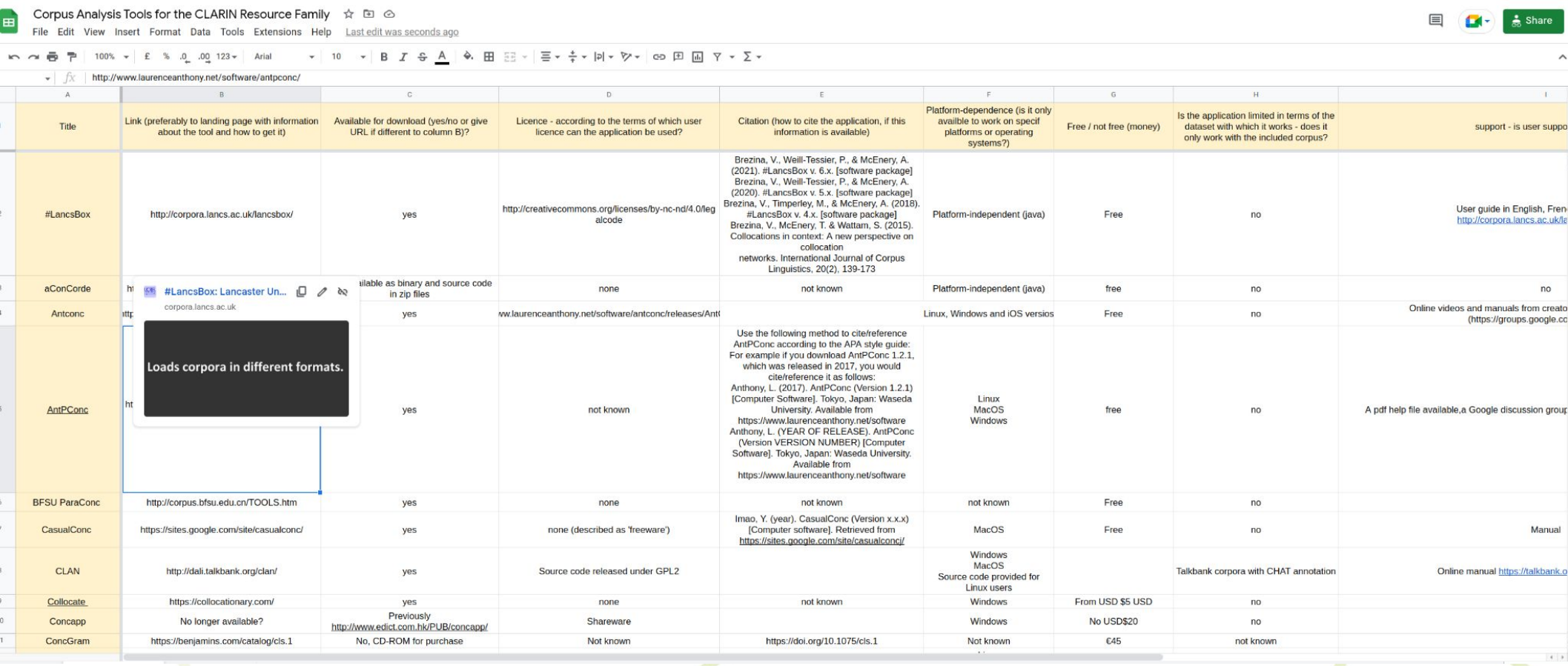
simple hic • 124,001
9,268.64 per million tokens • 0.93%

Get more space

KWIC

	Details	Left context	KWIC	Right context
1	poetry • uncert...	muta est.</s><s>sed nostram durare fidem duo pignora cogent: vel caper ille, nota frontem qui pingitur alba, vel levis	haec	et mobilibus circumdata bullis fistula, silvicolae munus memorabile Fauni.</s><s>sive caprum mavis vel Fauni ponere munus
2	poetry • uncert...	verbis?</s><s>iudicis e gremio victoris gloria surgat.</s><s>praeda mea est, quia Caesareas me dicere laudes mens iubet:	huic	semper debetur palma labori.</s><s>et mi sidereo cor movit Cynthius ore laudatamque chelyn iussit variare canendo.</s><s>pergite
3	poetry • uncert...	.</s><s>talis Phoebus erat, cum laetus caede draconis docta repercusso generavit carmina plectro: caelestes ulli si sunt,	hac	voce loquuntur!</s><s>venerat ad modulos doctarum turba sororum... huc huc, Pierides, volucri concedite saltu: hic
4	poetry • uncert...	sunt, hac voce loquuntur!</s><s>venerat ad modulos doctarum turba sororum... huc huc, Pierides, volucri concedite saltu:	hic	Heliconis opes florent, hic vester Apollo est!</s><s>tu quoque Troia sacros cineres ad sidera tolle atque Agamemnoniis opus
5	poetry • uncert...	ad modulos doctarum turba sororum... huc huc, Pierides, volucri concedite saltu: hic Heliconis opes florent,	hic	vester Apollo est!</s><s>tu quoque Troia sacros cineres ad sidera tolle atque Agamemnoniis opus hoc ostende Mycenis!</s><s>iam
6	poetry • uncert...	Heliconis opes florent, hic vester Apollo est!</s><s>tu quoque Troia sacros cineres ad sidera tolle atque Agamemnoniis opus	hoc	ostende Mycenis!</s><s>iam tanti cecidisse fuit!</s><s>gaudete, ruinae, et laudate rogos: vester vos tollit alumnus!</s><s>venerat en
7	poetry • uncert...	, pendet sacer hircus ab ulmo et iam nudatis cervicibus exuit exta.</s><s>ergo num dubio pugnant discrimine nati et negat	huic	aevo stolidum pecus aurea regna?</s><s>Saturni rediere dies Astraeaque virgo tutaque in antiquos redierunt saecula mores.</s>
8	prose • lulius ...	irascitur unda; mordent frena tigres, subeunt iuga saeva leones!</s><s>casta fave Lucina: tuus iam regnat Apollo!</s><s>Olim tibi	hos	libellos, Mavorti decus nostrum, me dicaturum esse promiseram, verum diu me inconstantia verecundiae retardavit et
9	prose • lulius ...	luminis vel augmenta vel damna, quantis etiam conversionibus maior ille quem ferunt perficeretur annus, qui quinque	has	stellas, Lunam etiam et Solem locis suis originibusque restituit, qui mille et quadringentorum et sexaginta et unius
10	prose • lulius ...	, quod quidam Atlanticum aiunt mare, terrae spatium in modum insulae undarum suarum circumfusione concludat.</s><s>	Haec	cum omnia mihi a te, Mavorti ornamentum bonorum, facili demonstrationis magisterio traderentur, ausus sum etiam
11	prose • lulius ...	disputationibus suis ex ipsis primo mathematicorum responsis quasi per cuniculum quendam ad labefactandam	hanc	totam scientiam conantur inrepere; qui quanto vehementius pugnant quantumque persuadendi vias quaerunt, tanto
12	prose • lulius ...	confirmant.</s><s>Neque enim esset eius vera substantia, nisi contra eam tantis argumentorum viribus niterentur.</s><s>Nec	hoc	est admiratione dignum, cum sciamus inter ipsos quanta sit de ipsa deorum natura dissensio quantisque disputationum
13	prose • lulius ...	, sed nihil procurare definiant, alii et esse et rerum nostrarum curam procurationemque suscipere, et tanta sint	hi	omnes in varietate et dissensione versati, ut longum et alienum sit, hoc praesertim tempore cum aliud opus adgressi
14	prose • lulius ...	procurationemque suscipere, et tanta sint hi omnes in varietate et dissensione versati, ut longum et alienum sit,	hoc	praesertim tempore cum aliud opus adgressi sumus, singulorum enumerare sententias.</s><s>Nam alii et figuras his pro
15	prose • lulius ...	sit, hoc praesertim tempore cum aliud opus adgressi sumus, singulorum enumerare sententias.</s><s>Nam alii et figuras	his	pro arbitrio suo tribuunt et loca adsignant, sedes etiam constituunt et multa de actibus eorum vitaque describunt et
16	prose • lulius ...	bono ac malo diversae inter se mobilesque sententiae.</s><s>Quarum explicationem nunc praetermittendam puto; neque enim	hoc	genus disputationis intravimus nec ad haec refutanda vel confirmanda animum nostrum consiliumque formavimus, sed
17	prose • lulius ...	sententiae.</s><s>Quarum explicationem nunc praetermittendam puto; neque enim hoc genus disputationis intravimus nec ad	haec	refutanda vel confirmanda animum nostrum consiliumque formavimus, sed omnia haec ideo brevi oratione
18	prose • lulius ...	disputationis intravimus nec ad haec refutanda vel confirmanda animum nostrum consiliumque formavimus, sed omnia	haec	ideo brevi oratione perstrinximus, ut contradicentium contentiosa commenta cunctis hominibus patefacto sui
19	prose • lulius ...	securus stellarum iudiciis adscribat.</s><s>" Sit iniquus, sit perfidus, sit malivolus, sit inreligiosus, quia	haec	vitia mala Mercurii cum Marte perfecit constellatio!</s><s>Cur animum nostrum fortitudinis studio roboramus, cur erecta
20	prose • lulius ...	bonis radiatonibus temperatus ad desideria nos et cupiditates istius virtutis inflammat.</s><s>" Etiam singulos homines	hac	disputationis oratione conveniunt:" Frustra habes, o bone vir, dilectum bonarum rerum ac malarum fugiendi

Rows per page: 20 1–20 of 124,001 1 / 6,201



Corpus Analysis Tools for the CLARIN Resource Family

File Edit View Insert Format Data Tools Extensions Help Last edit was seconds ago

Share

http://www.laurenceanthony.net/software/antpconc/

Title	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 availble to work on specif platforms or operating systems?)	Free / not free (money)	Is the application limited in terms of the dataset with which it works - does it only work with the included corpus?	support - is user suppo
#LancsBox	http://corpora.lancs.ac.uk/lancsbox/	yes	http://creativecommons.org/licenses/by-nc-nd/4.0/legalcode	Brezina, V., Weill-Tessier, P., & McEnery, A. (2021). #LancsBox v. 6.x. [software package] Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). #LancsBox v. 5.x. [software package] Brezina, V., Timperley, M., & McEnery, A. (2018). #LancsBox v. 4.x. [software package] Brezina, V., McEnery, T. & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. International Journal of Corpus Linguistics, 20(2), 139-173	Platform-independent (java)	Free	no	User guide in English, Fren http://corpora.lancs.ac.uk/la
aConCorde	ht	ilable as binary and source code in zip files	none	not known	Platform-independent (java)	free	no	no
Antconc	http	yes	ww.laurenceanthony.net/software/antconc/releases/Ant(		Linux, Windows and iOS versios	Free	no	Online videos and manuals from creato (https://groups.google.co
AntPConc	ht	yes	not known	Use the following method to cite/reference AntPConc according to the APA style guide: For example if you download AntPConc 1.2.1, which was released in 2017, you would cite/reference it as follows: Anthony, L. (2017). AntPConc (Version 1.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software Anthony, L. (YEAR OF RELEASE). AntPConc (Version VERSION NUMBER) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software	Linux MacOS Windows	free	no	A pdf help file available,a Google discussion group
BFSU ParaConc	http://corpus.bfsu.edu.cn/TOOLS.htm	yes	none	not known	not known	Free	no	
CasualConc	https://sites.google.com/site/casualconc/	yes	none (described as 'freeware')	lmao, Y. (year). CasualConc (Version x.x.x) [Computer software]. Retrieved from https://sites.google.com/site/casualconc/	MacOS	Free	no	Manual
CLAN	http://dali.talkbank.org/clan/	yes	Source code released under GPL2		Windows MacOS Source code provided for Linux users		Talkbank corpora with CHAT annotation	Online manual https://talkbank.o
Collocate	https://collocationary.com/	yes	none	not known	Windows	From USD $5 USD	no	
Concapp	No longer available?	Previously http://www.edict.com.hk/PUB/concapp/	Shareware		Windows	No USD$20	no	
ConcGram	https://benjamins.com/catalog/cls.1	No, CD-ROM for purchase	Not known	https://doi.org/10.1075/cls.1	Not known	€45	not known	

#LancsBox: Lancaster Un...
corpora.lancs.ac.uk

Loads corpora in different formats.

Please contribute by adding comments to the spreadsheet at:

https://bit.ly/CLARINCorpusTools

Corpus Analysis Tools for the CLARIN Resource Family

File Edit View Insert Format Data Tools Extensions Help Last edit was seconds ago

Citation (how to cite the application, if this information is available)

Title	Link	Country (where website is based, if relevant)	Free / not free (money)	Is there a specific user licence governing the use of the online tools (state name and provide link if yes)?	Citation (how to cite the application, if this information is available)	Registration required - do users need to register to use the tool?	Shibboleth authentication? (Can users log in with their own institutional single-sign on?)	Languages - if the application is restricted to a fixed set of installed copora, what langauges are avaialble? (use ISO 639-2 codes or plain text names)	Can you upload your own corpus?	support (may have support as part of a project but time imited so not sustainable), say YES if eg an email address given	Description	Other
COSMAS 2	https://cosmas2.ids-mannheim.de/cosmas2-web/	Germany	Free									
COSMAS II	https://cosmas2.ids-mannheim.de/cosmas2-web	Germany	Free								the German reference corpus - DeReKo, historical corpora and others.	
CQPweb (Lancaster)	https://cqpweb.lancs.ac.uk/	UK	Free	no		yes	no (certain authorizations granted to users with email in the .ac.uk domain)	eng, ara, fre, ita, nor, pol, lat Numerous South-Asian languages	yes	Online help system, bug reporting system, email address for queries	Will be superseded by Zigguratt	
CQPweb (Beijing Foreign Studies University)	CQPweb cqpweb.lancs.ac.uk	China	Free			Yes - no indication of how obtain credentials in order to authenticate						
Croatian National Corpus	http://filip.ffzg.hr/	Croatia	Free			no		hrv			NoSketchEngine interface to the Croatian National Corpus	
DWDS	https://www.dwds.de/	Germany	Free			yes	no	deu	It works with a range of corpora available after registration: chat, law, German newspapers, blogs, historical texts		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	

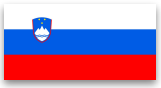
Please contribute by adding comments to the spreadsheet at:

https://bit.ly/CLARINCorpusTools

CLARIN corpus resources and tools

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

Online interfaces:

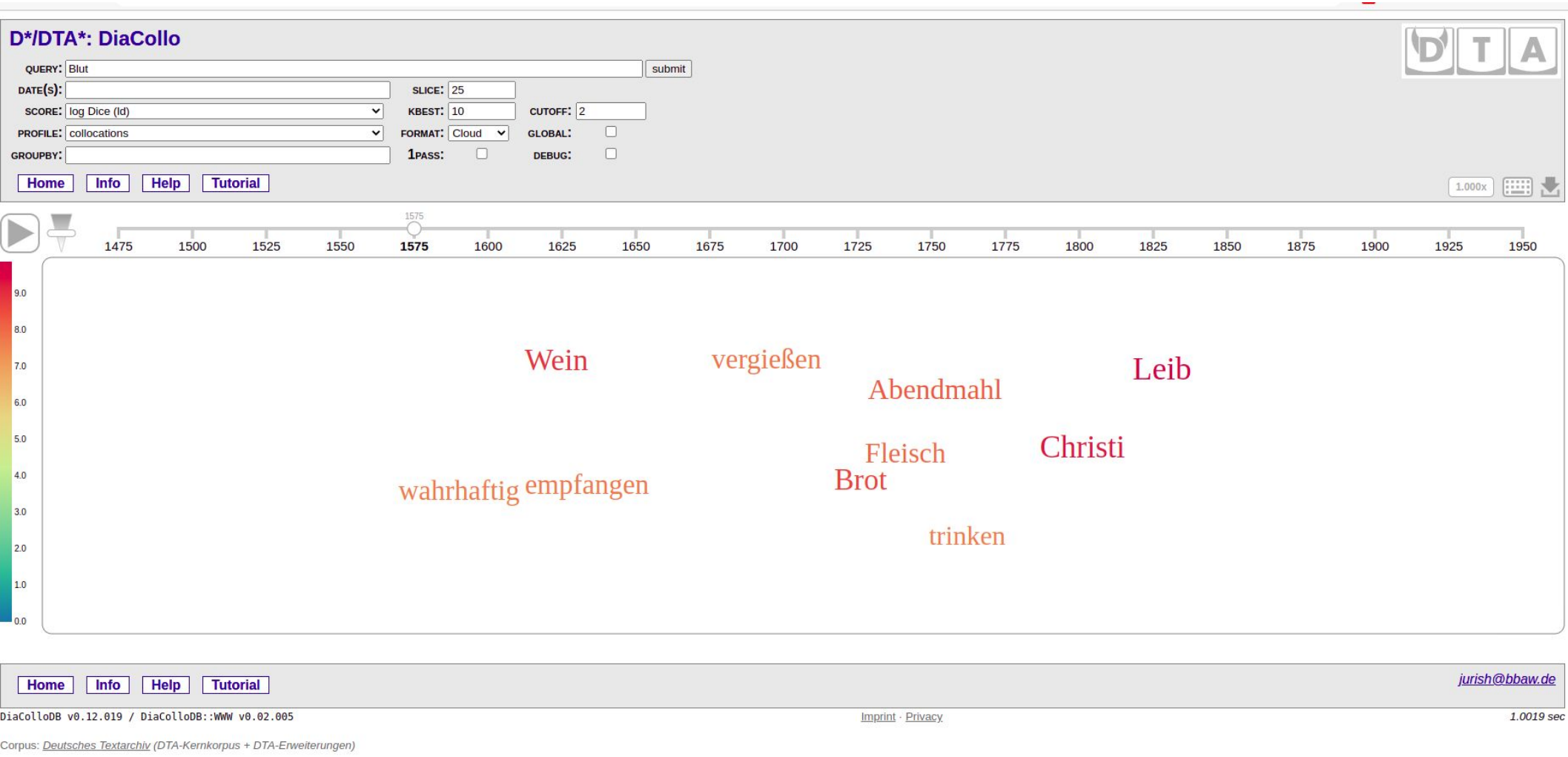
- Corpuscle
- Korp
- KonText
- NoSketch Engine 
- D* (Diacollo demo)
- TEITOK

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 | NEAR(($l=@'Blut') =1, ($l=@'Wein' WITH $p=@'NN') =2, 8) #SEPARATE #asc_date[1575-00-00,1599-9 | submit | + | CabErr

1:	[dtae:magirus_widerlegung_1580:8]	... Altars / daß daselbst warhafftig der Leib vnnd	**Blut**[1]	im Brot vnd **Wein**[2] / werde mündtlich geessen...
2:	[dtae:magirus_widerlegung_1580:18]	... Brot vnd **Wein**[2] / seinen warhafftigen Leib vnd	**Blut**[1]	/ zuessen vñ zutrincken darreiche vnd gebe:...
3:	[dtae:magirus_widerlegung_1580:18]	... die vngläubige vnd vnwürdige / den Leib vñ	**Blut**[1]	Christi mit Brot vnd **Wein**[2] / im heiligen...
4:	[dtae:magirus_widerlegung_1580:19]	... **Wein**[2] / aber gar nicht der Leib vnd	**Blut**[1]	Christi / gegenwertig dargereicht vñ empfangen werde /...
5:	[dtae:magirus_widerlegung_1580:31]	... wir vns (ob der ware Leib vnd	**Blut**[1]	Christi leiblich im Brot vnd **Wein**[2] sey)...
6:	[dtae:magirus_widerlegung_1580:31]	... anders gewölt / dann daß Christi Leib vnd	**Blut**[1]	war hafftig im Nachtmal mit Brot vnd **Wein**[2]...
7:	[dtae:magirus_widerlegung_1580:35]	... / von einer wesentlichen Gegenwertigkeit des Leibs vnd	**Bluts**[1]	Christi / vnder gestalt des Brots vnd **Weins**[2]...
8:	[dtae:magirus_widerlegung_1580:36]	... Brots vnd **Weins**[2] / in den Leib vnnd	**Blut**[1]	Christi (also daß kein Brot vnd Wein...
9:	[dtae:magirus_widerlegung_1580:36]	... Brots vnd Weins / in den Leib vnnd	**Blut**[1]	Christi (also daß kein Brot vnd **Wein**[2]...
10:	[dtae:magirus_widerlegung_1580:37]	... / vnd mit dem sichtbarn **Wein**[2] / das	**Blut**[1]	Christi dargereicht werde.

~HTML | ~Hist | Home | Query Lizard | Previous | Next | Help | NEAR(($l=@'Blut') =1, ($l=@'Wein' WITH $p=@'NN') =2, 8) #SEPARATE #asc_date[1575-00-00,1599-9 | submit | + | CabErr

jurish@bbaw.de

D* OpenSearch API version 0.60

0.693405 sec

Corpus: Deutsches Textarchiv (DTA-Kernkorpus + DTA-Erweiterungen)

Anmelden (DTAQ) DWDS dlexDB CLARIN-D

D T A

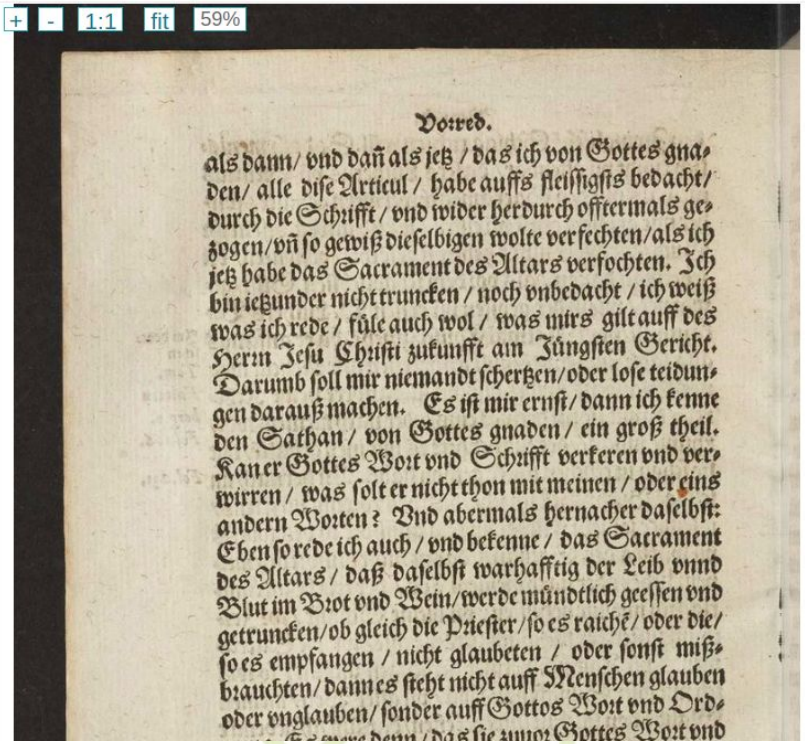
suchen Hilfe Texte ▼ Projekt ▼ Dokumentation ▼ Impressum

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 << vorherige Seite nächste Seite >>

+ - 1:1 fit 59%



Vorred.

als dann/ vnd dañ als jetz / das ich von Gottes gnaden/ alle dise Articul / habe auffs 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 schertzen/oder lose teidungen darauß machen. Es ist mir ernst/ dann ich kenne den Sathan / von Gottes gnaden / ein groß theil. Kan er 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 vnnd 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 Gottos Wort vnd Ord-

+ -

als dann / vnd dañ als jetz / das ich von Gottes gnaden / alle dise Articul / habe auffs 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 schertzen / 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 vnnd 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 Gottos 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 vnnd verendert.

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 availble to work on specif 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!

- Spot gaps, inaccuracies, missing information in https://bit.ly/CLARINCorpusTools;
- 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

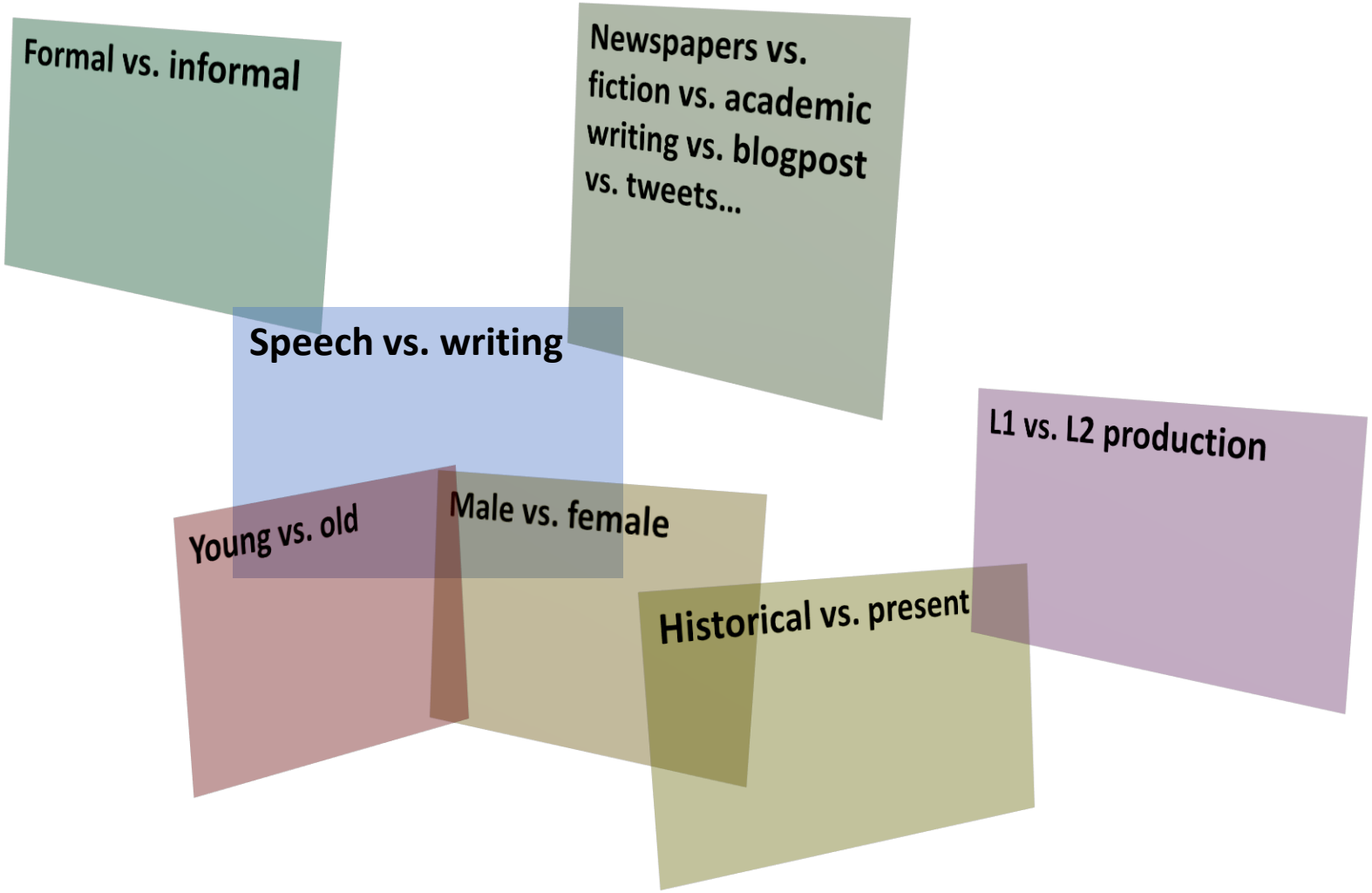
Lancaster University

@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.

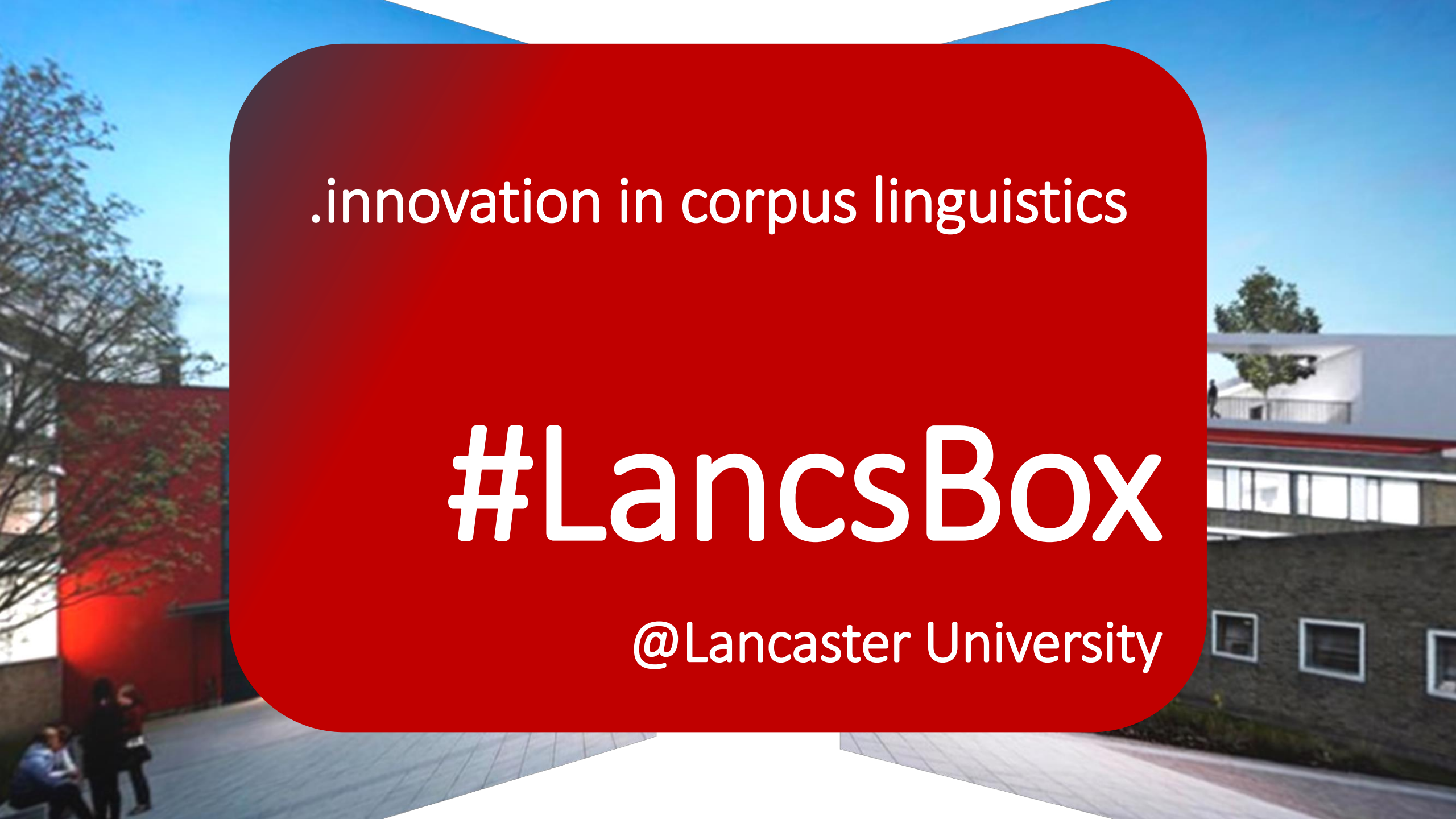
Multidimensional data



Lancaster
University

Lanca
Univer

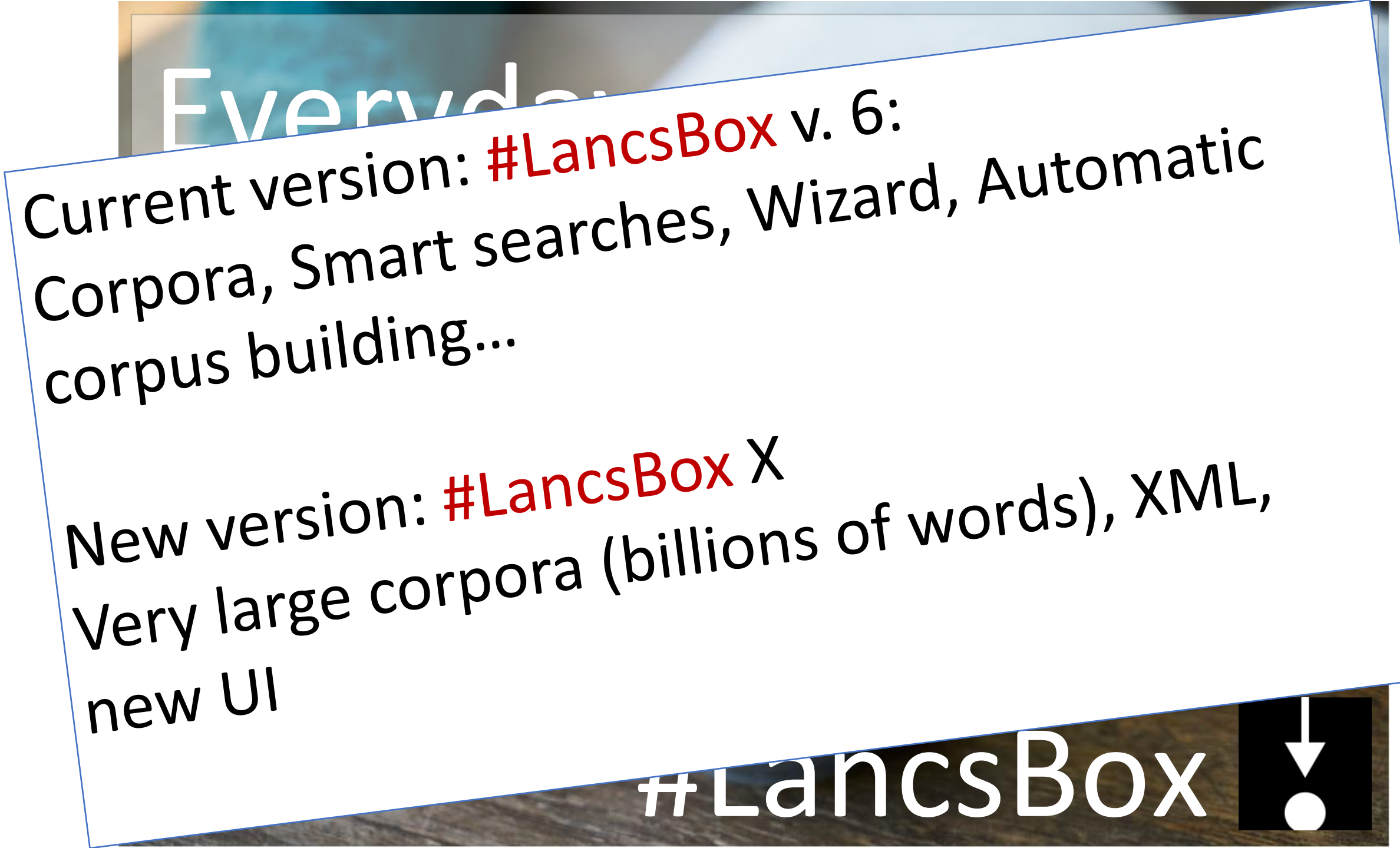
ster
sity



.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

futurelearn.com/courses/corpus-linguistics

Update

Lancaster University

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!

Pin Like Reply Bookmark Report

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!

Pin Like Reply Bookmark Report

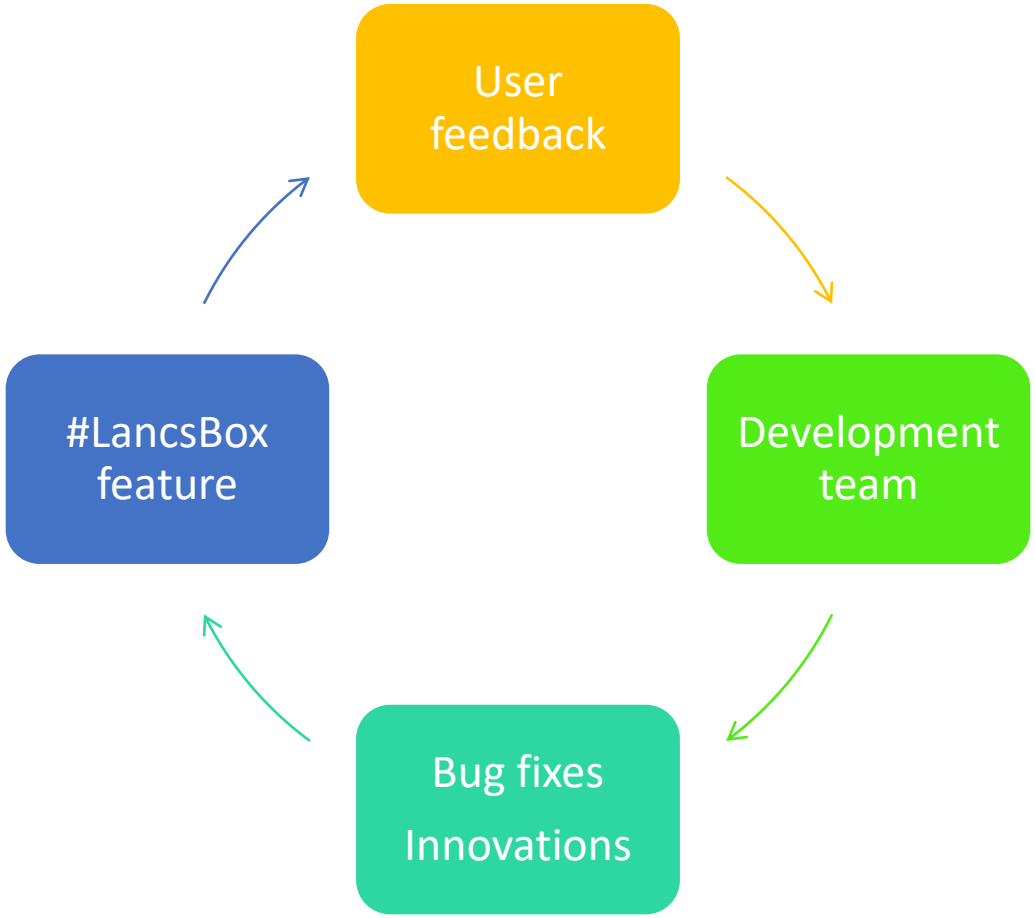
BH

Becky H Follow 21 NOV

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

Pin Like Reply Bookmark Report

Feedback cycle



Download #LancsBox:
http://corpora.lancs.ac.uk/lancsbox/

Type: some, Corpus: Corpus 1

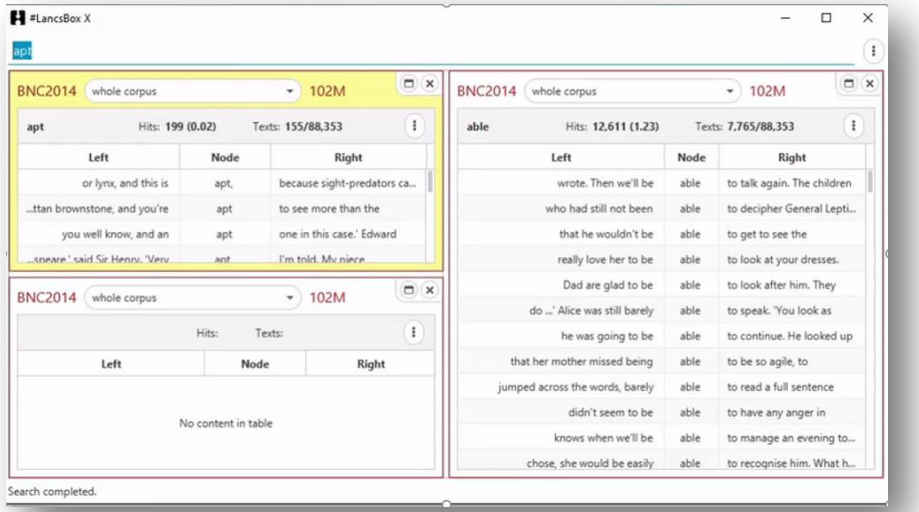
Showing 1617 occurrences in window 5L, 5R.

File	L5	L4	L3	L2	L1	Node	R1	R2	R3	R4	R5
BROWN_F.txt	to	have	believe	me	—	some	see	only	darkness	the	bitter
BROWN_H.txt	projects	requiring	special	maturity	and	some	open	only	to	men	or
BROWN_J.txt	self	While	some	suppression	and	some	denial	are	not	only	necessary
BROWN_D.txt	be	only	a	punishment	for	some	termination	seems	necessary	in	a
BROWN_L.txt	sobered	up	enough	to	get	some	sense	Maybe	he	only	intended
BROWN_E.txt	to	green	however	for	in	some	situations	only	a	yellow	indication
BROWN_J.txt	de	jure	population	or	in	some	districts	only	the	de	facto
BROWN_J.txt	question	of	epidemic	disease	merits	some	discussion	Only	a	limited	effort
BROWN_L.txt	expense	money	We	ll	need	some	at	least	if	only	bus
BROWN_G.txt	means	only	the	presence	of	some	minor	characteristic	not	quite	usual
BROWN_J.txt	has	to	go	back	only	some	sixty	years	The	first	mention
BROWN_J.txt	end	where	of	Costaggini	only	some	foliage	has	been	washed	at
BROWN_H.txt	that	he	now	devoted	only	some	6-1/2	hours	per		
BROWN_J.txt	such	suggestion	accounts	for	only	some	of	the	difficulti		
BROWN_K.txt	paused	for	an	instant	only	some	of	them	glancing		
BROWN_B.txt	pro-Comm...	rebel	force	numbers	only	some	2,000	tough	Viet		
BROWN_B.txt	initialed	miscellany	and	with	only	some	2,000	subscribers	Little		
BROWN_J.txt	given	only	the	information	that	some	people	have	somethi		
BROWN_G.txt	a	constant	stream	of	visitors	some	stopping	only	to		

Filter: only | Any text field | Case sensitive

Save

Version 1.0 (2015)



Version X (2022)

#LancsBox 6.0

KWIC | GraphColl | Whelk | Words | Ngrams | Text | Wizard

Corpora | KWIC: some

Search

Search some **Occurrences** 7,969 (15.86) Texts 13 **▼ Corpus** BNC2014-baby **▼ Context** 7 **▼ Display Text**

Index	File	Left	Node	Right
1	Academic_bo	beastly; creatures of dubious or shifting classification.	Some	of these have been the subject of
2	Academic_bo	fictional world with its own laws... Moreover,	some	kinds of metamorphosis play a crucial part
3	Academic_bo	this might have been the case for	some	contemporary readers, who cannot fail to have
4	Academic_bo	others. There were overtly political beasts, too:	'Some	of the political implications of Nietzsche's views
5	Academic_bo	Übermensch, "Superman", or "blond beast" occur with	some	frequency from the mid- 1890s; and there
6	Academic_bo	with the search for identity, or in	some	cases its antithesis, the refusal to develop'.20
7	Academic_bo	fears from other decades and there were	some	in society who felt no such insecurity.
8	Academic_bo	study invariably happens within, or itself constitutes,	some	kind of journey. Travel functions as a
9	Academic_bo	In his introduction to an anthology of	some	of the important works of social exploration
10	Academic_bo	but as historically inflected, an outline of	some	of the events from 1885 to 1900,
11	Academic_bo	one historical survey of the period: Without	some	fresh influx of wealth the nobility as
12	Academic_bo	Enlightenment, or Romantics,102 even if there is	some	continuity between them. From Darwin we see
13	Academic_bo	Most obviously, Darwinian thought was taken by	some	to reduce the possibility of human agency
14	Academic_bo	as working on the past to create	some	new present'. In 'ideas of divination' and
15	Academic_bo	what it is like to be in	some	other contemporary condition: bereaved, unemployed, insane'.115 Admittedly,
16	Academic_bo	episodes, and fi nally, to arcs, with	some	consideration of sequences. Beats Th e smallest
17	Academic_bo	and forth between these main characters, and	some	minor characters as well. Transitions between these
18	Academic_bo	diff erent ways in diff erent shows:	some	shows always end on a cliffh anger,
19	Academic_bo	'previously on...' clip before each episode. But	some	shows, like Th e Good Wife(CBS,
20	Academic_bo	(or futures)...'. 96 At the other extreme,	some	episodes introduce characters only to quickly dispose
21	Academic_bo	in the serial's narrative, they usually leave	some	problems unresolved, or 'dangling', 106 which subsequent
22	Academic_bo	from a few beats to several episodes-	some	can last for the whole narrative. Newman
23	Academic_bo	of approaches of this kind, and outlined	some	interpretative justifications for such a reading of
24	Academic_bo	relating deeply to twenty- first- century poets.7	Some	of these areas or key thoughts animate
25	Academic_bo	the section 'Grace and Surprises', Ford explores	some	broader theological points raised by Dante's poem
26	Academic_bo	planned it this way ab initio in	some	imagined blueprint of his poetic afterlife. I
27	Academic_bo	each of Apichatpong's films is offset by	some	other story: a fugitive buffalo returns to
28	Academic_bo	annotations (no longer seen, at least by	some,	as defacement but as adding charm) which
29	Academic_bo	destructive and often controversial. Conservation documentation goes	some	way to preserving evidence through recording, but
30	Academic_bo	'father' of modern scientific conservation) and, with	some	modification, has continued in use (Gilberg 1987;
31	Academic_bo	al. 2002). It is still used for	some	tablets, although a more cautious, and less
32	Academic_bo	of damaged and 'fraying' documents. By sacrificing	some	of the material (the edges) she aimed
33	Academic_bo	been in use by conser- vators for	some	time to detect obscured surface detail, such
34	Academic_bo	self- censorship and eased competition with rivals.	Some	however tried to get round the new

Filtering complete

Version 6.0 (2021)

#LancsBox highlights

1. Smart searching

GRAMMATICAL ST

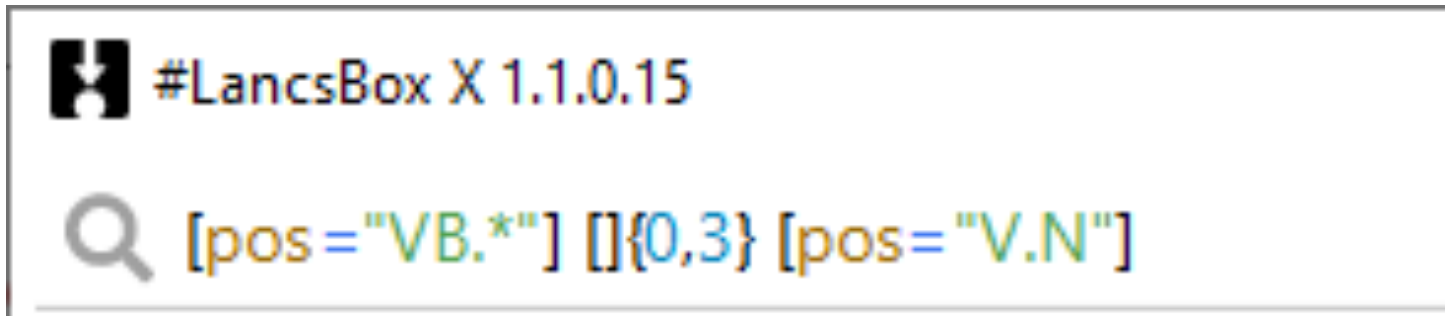
Smart searches

- 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

#LancsBox X 1.1.0.15

cat

BNC2014 whole corpus 102M

dog Hits: 9,402 (91.80) Texts: 2,988/88,171

File	Left	Node	Right
MagRe	George stares at protective pet	dog	Lupo - too cute! There are
NewRei-final			H ALAN DAVIES Tu
AcaSoc			or
AcaHu			d to be
AcaHu			lowels in
NewM			he Ronnie
Sp2m1			d we ignore
NewRe	business. What happened to the	dog	that ate nothing but garlic?
AcaHu	satyr, a shepherd with a	dog,	a slave and the four
FictMis	he knew Sydney. And the	dog -	even if it had come
FictMis	time, after four or five	dog	lifetimes - would not have come
FictMis	he follows Sydney and his	dog	out of the kitchen, Lewis
FictMis	to anyone's face. Your	dog,	he would write, is too
FictMis	gone in his seat. The	dog	is standing on the back
FictMis	's ear. There is hot	dog	breath on Lewis's neck,
FictMis	from the tip of the	dog	's huge tongue. They are
Sp1m1	out another bike then honestly	dog	eat dog what really annoys
Sp1m1	bike then honestly dog eat	dog	what really annoys me oh
MagT3	markets, Nokia is still top	dog -	or at least top dog
MagT3	dog - or at least top	dog	in terms of what they

new family photo with Prince George The Duke and Duchess of Cambridge have marked her first Mother's Day as a mummy by releasing a brand new family portrait. In the photo, William and Kate can be seen smiling at the camera while adorable Prince George stares at protective pet **dog,** Lupo - too cute! There are no signs of any sleepless nights on the happy faces of Will and Kate, while baby George has certainly grown since his last appearance in an official photograph. We last saw the world's most famous baby in his official Christening photos in October 2013.

BNC2014 whole corpus ... –

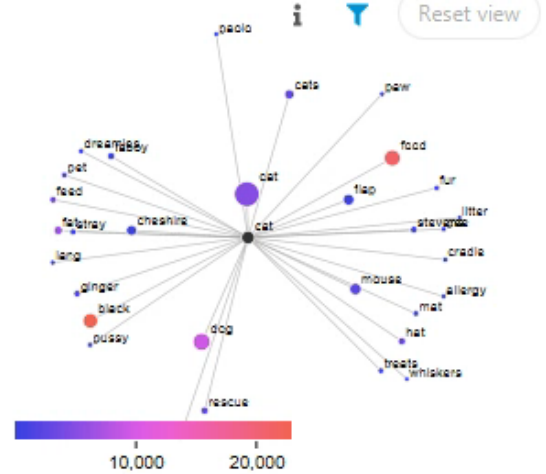
cat ×

cat Hits: 4,928 (48.12)

"Log Dic...

Collocate	Distribu...	Freq.	Freq. (c...
whiskers		10	155
stray		18	541
dog		120	9,402
paolo		11	169
black		89	22,838
fat		31	6,250

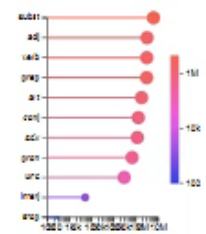
Reset view



BNC2014 academic prose 20M class

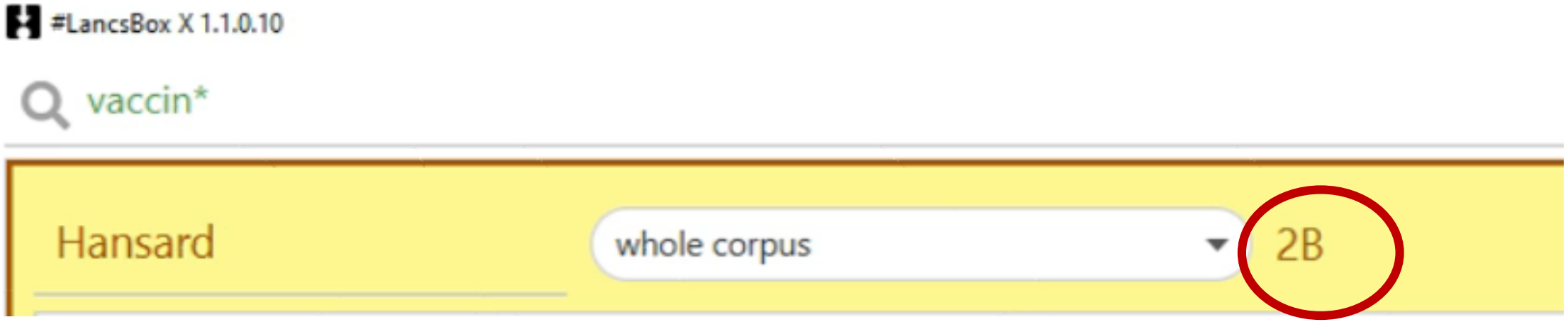
Terms: 11

Term	Freq...	Rel. ...	ARF	Range	Ran...	CV	Juill...	DF
subst	5,8...	29...	4,3...	2,8...	10...	0.11	1.00	
adj	3,0...	15...	2,0...	2,8...	10...	0.15	1.00	
verb	2,9...	14...	1,9...	2,8...	10...	0.13	1.00	
prep	2,9...	14...	2,1...	2,8...	10...	0.09	1.00	
art	1,7...	86,...	1,1...	2,8...	10...	0.24	1.00	
conj	1,2...	61,...	82...	2,8...	10...	0.17	1.00	
adv	1,0...	55,...	63...	2,8...	99....	0.28	0.99	



Searching for cat.

3. #LancsBox X: very large corpora



#LancsBox X 1.1.0.10

vaccin*

Hansard | whole corpus | 2B

vaccin* Hits: 75,771 (40.06) Texts: 8,180/75,317

66M "Text: date"=dateRan... 26,790 (406.35) 946/2,050

File	Left	Node	Right	Text: da
LordChan	everyone has access to healthcare. We will invest more in	vaccines	and research, so that developing countries can benefit	2020-01-07
LordChan	The same is true for GAVI, as the efficacy of	vaccines	is massively reduced if the recipient is malnourished. W	2020-01-07
LordChan	to spend large sums of money on girls' education,	vaccinations	or climate change, let us not just tick the box,	2020-01-07
LordChan	rnment on their work in these areas. Education, climate and	vaccines	are key building blocks for creating a world in which	2020-01-07
LordChan	it is also why we have run a vigorous flu	vaccination	programme to prevent people from getting into that p	2020-01-09
LordChan	clinical trial into scurvy, to proton beam therapy and mass	vaccination	programmes. The Medicines and Medical Devices Bill w	2020-01-09
CommCh	children in this country are not protected by the MMR	vaccine	when they start school? May we have an urgent debate	2020-01-09
CommCh	doing what they can to increase the take-up of measles	vaccines	to ensure that as many children as possible are covered	2020-01-09
CommCh	have a great responsibility to ensure that their children are	vaccinated.	Let me add, without going into too much detail, that	2020-01-09
CommCh	but public health depends on people's use of the	vaccines	that are available to them, and that should be encoura	2020-01-09
CommCh	also means supporting the development and deployment of	vaccines;	an important signal of our commitment will be when w	2020-01-13

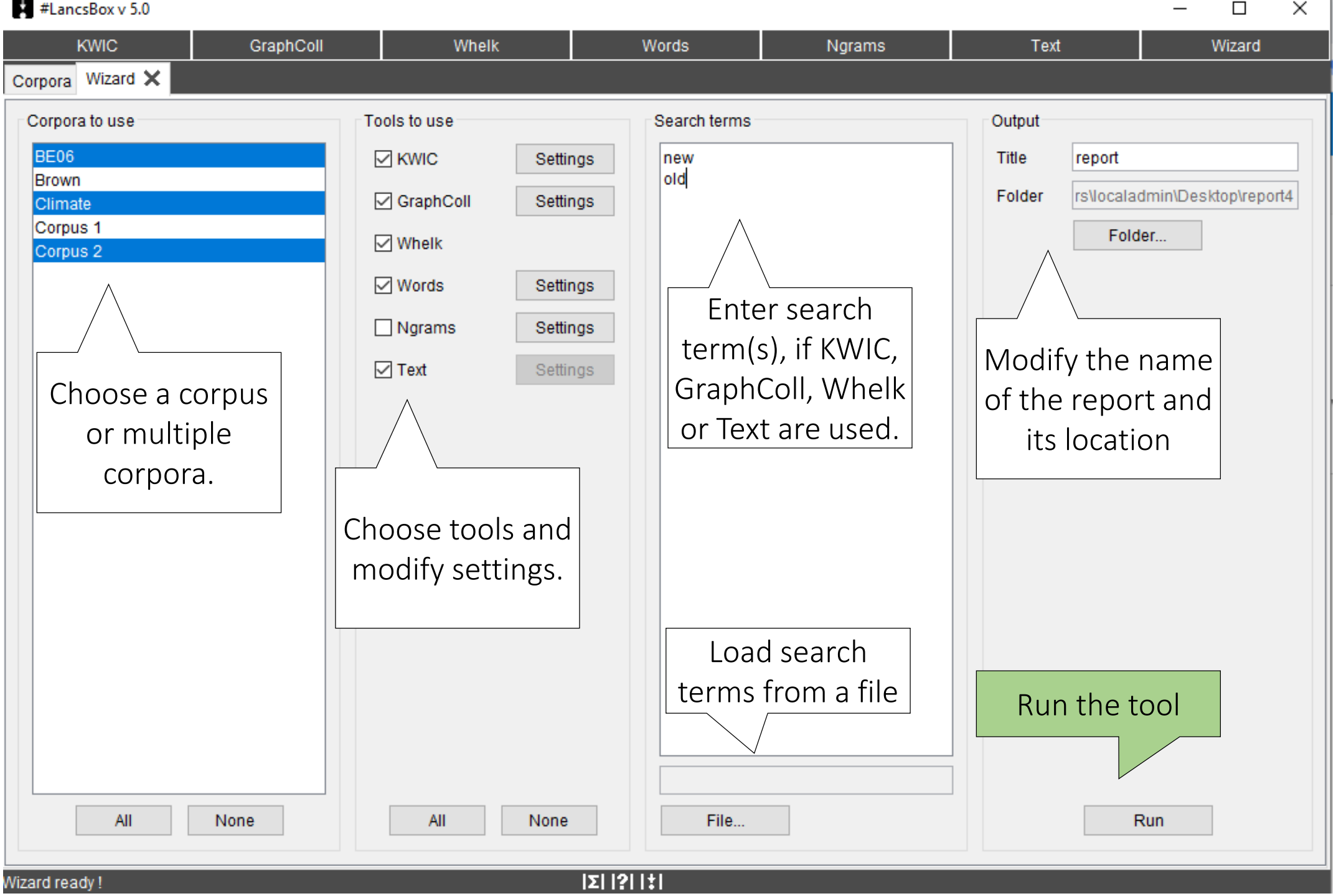
Search completed.

#LancsBox X 1.1.0.10

vaccin*

Hansard | whole corpus | 2B | word | L5 | – | R5

Reset view


#LancsBox v 5.0
KWIC
GraphColl
Whelk
Words
Ngrams
Text
Wizard
Corpora
Wizard
Corpora to use
BE06
Brown
Climate
Corpus 1
Corpus 2
Choose a corpus or multiple corpora.
All
None
Tools to use
KWIC
Settings
GraphColl
Settings
Whelk
Words
Settings
Ngrams
Settings
Text
Settings
Choose tools and modify settings.
All
None
Search terms
new
old
Enter search term(s), if KWIC, GraphColl, Whelk or Text are used.
Load search terms from a file
File...
Output
Title
report
Folder
rs\localadmin\Desktop\report4
Folder...
Modify the name of the report and its location
Run the tool
Run
Wizard ready !

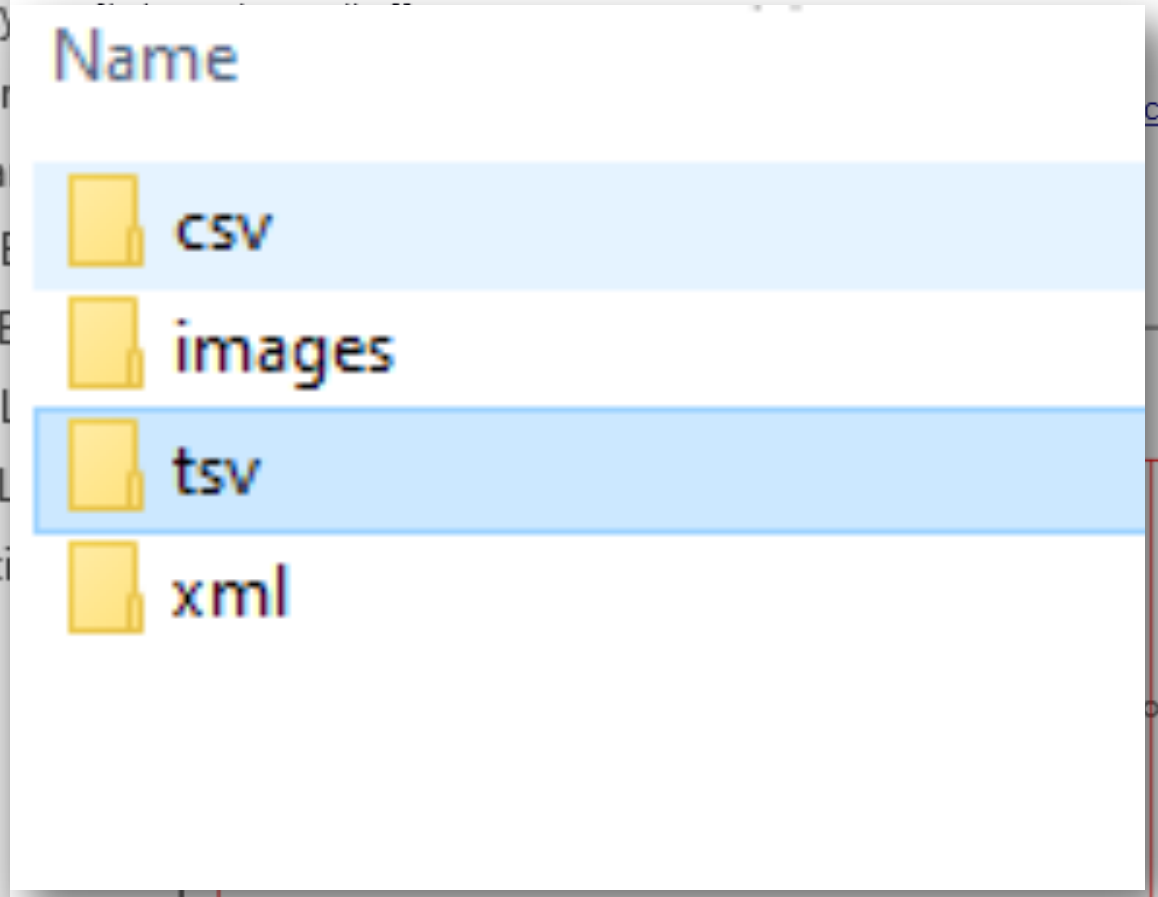
1 Introduction
2 Method
2.1 Data
2.2 Procedure
3 Results
3.1 General overview: Frequency
3.2 Corpus comparison: Keywor
3.3 Specific searches: Concorda
3.3.1 Search term "good" in E
3.3.2 Search term "great" in E
3.3.3 Search term "good" in L
3.3.4 Search term "great" in L
3.4 Word associations: Collocati
3.5 Statistical analysis
4 References
5 Appendix

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,

csv.



Advantages of #LancsBox

- More *time* for exploration.
- More *time* for multiple perspectives.
- More *time* for linguistic interpretation.
- More *time* for things that matter.

for
god
thank
much
goodness
you

#LancsBox: Lancaster University corpus toolbox

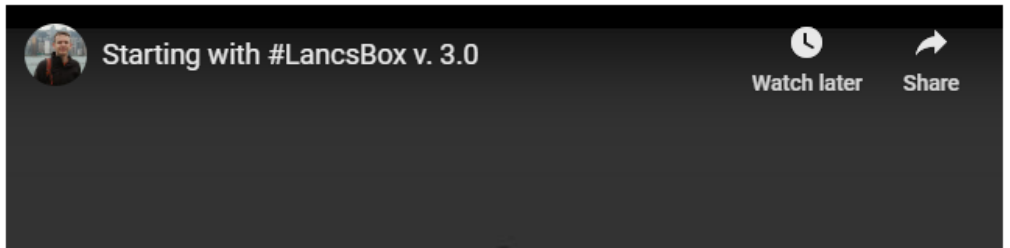
User guide [pdf]

Guide de l'utilisateur [pdf]

Overview and data

corpora.lancs.ac.uk/lancsbox

Starting with #LancsBox [pdf]



AntConc:

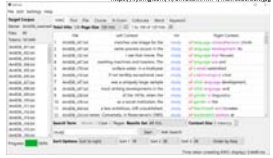
A corpus analysis toolkit for concordancing and text analysis

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

Clarin Symposium
September 15, 2022 (online)

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

AntLab tools; Release history; User statistics

History of AntConc:

One of the 17 "AntLab" freeware language processing tools

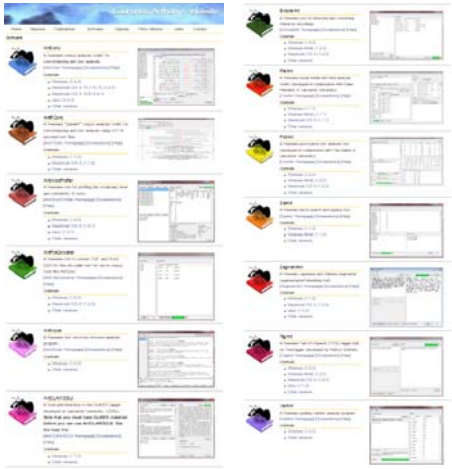
Data Collection Tools

- AntCorGen
- AntFileConverter
- AntFileSplitter
- FireAnt

AntLab Tools
www.laurenceanthony.net/software

History of AntConc:

One of the 17 "AntLab" freeware language processing tools



Data Cleaning Tools

- EncodeAnt
- SarAnt
- SegmentAnt

AntLab Tools

www.laurenceanthony.net/software

History of AntConc:

One of the 17 "AntLab" freeware language processing tools

Data Annotation Tools

- AntMover
- CLAWSAnt
- SegmentAnt
- TagAnt

AntLab Tools

www.laurenceanthony.net/software

History of AntConc:

One of the 17 "AntLab" freeware language processing tools

Data Analysis Tools

- AntConc
- AntPConc
- AntWordProfiler
- FireAnt
- ProtAnt
- VariAnt

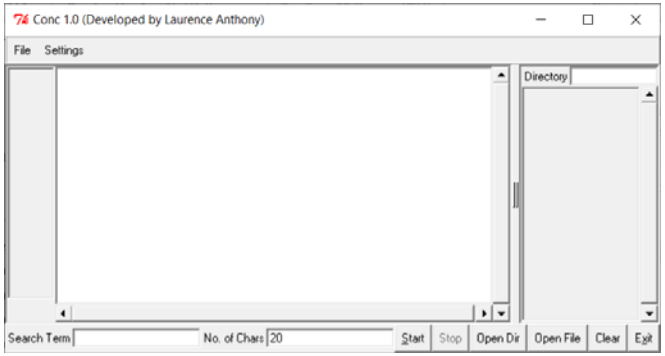
AntLab Tools

www.laurenceanthony.net/software

History of AntConc:

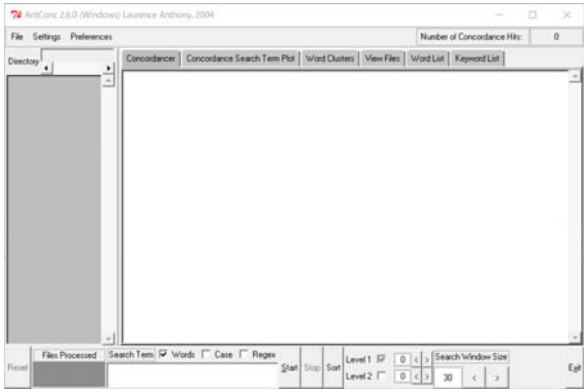
Four major releases

- Version 1.0-1.1: Simple concordancer for Windows and Linux
  - tools: concordance
  - language: Perl + Tk



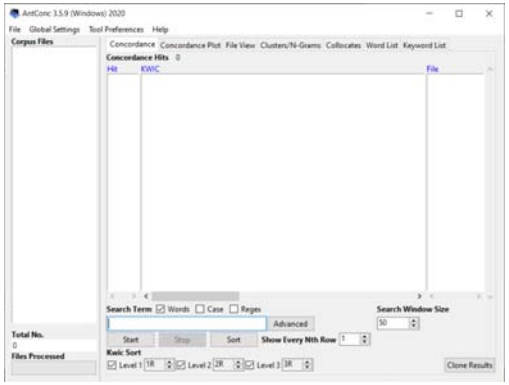
History of AntConc:
Four major releases

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



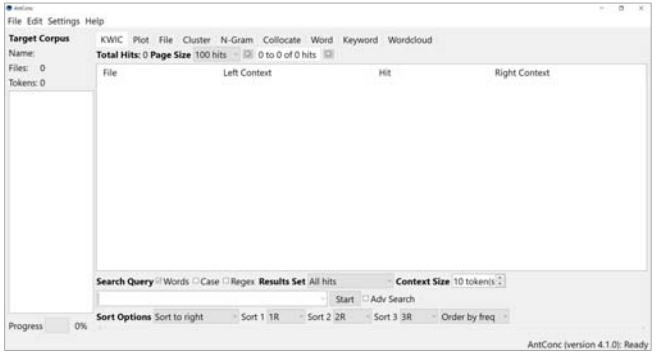
History of AntConc:
Four major releases

- 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

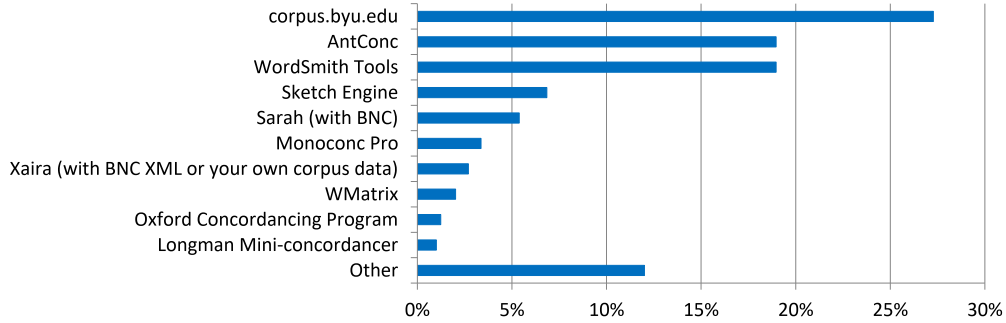


History of AntConc:
Four major releases

- 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



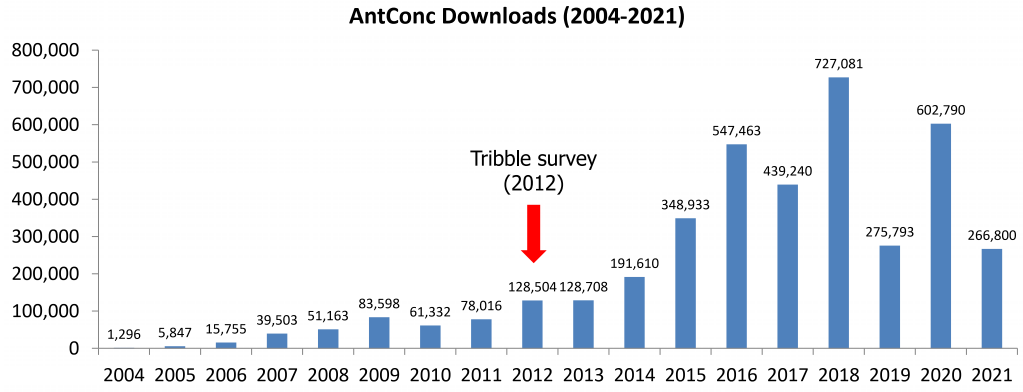
History of AntConc:
User statistics



"Which computer programs do you use for analysing corpora?"
International survey of corpus linguists. Reponses: 891. (Tribble, 2012)

History of AntConc:

User statistics



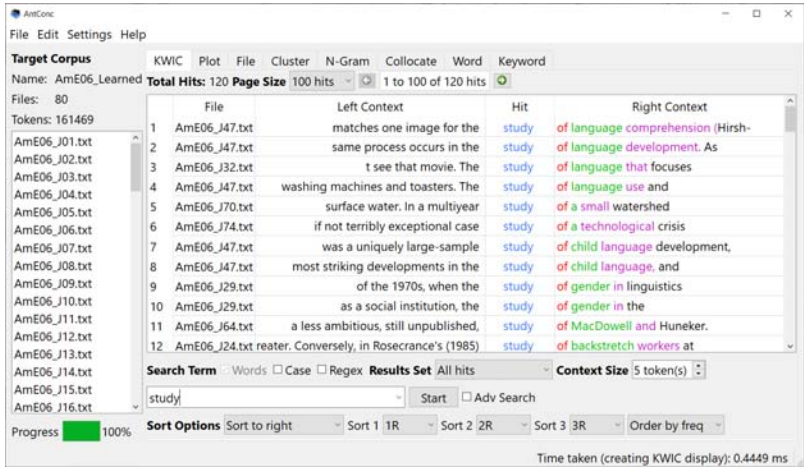
used in over 140 countries; YouTube tutorial videos viewed over 500,000 times

Overview of AntConc 4

design concept; tools and features

Overview of AntConc 4:

Design concept



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

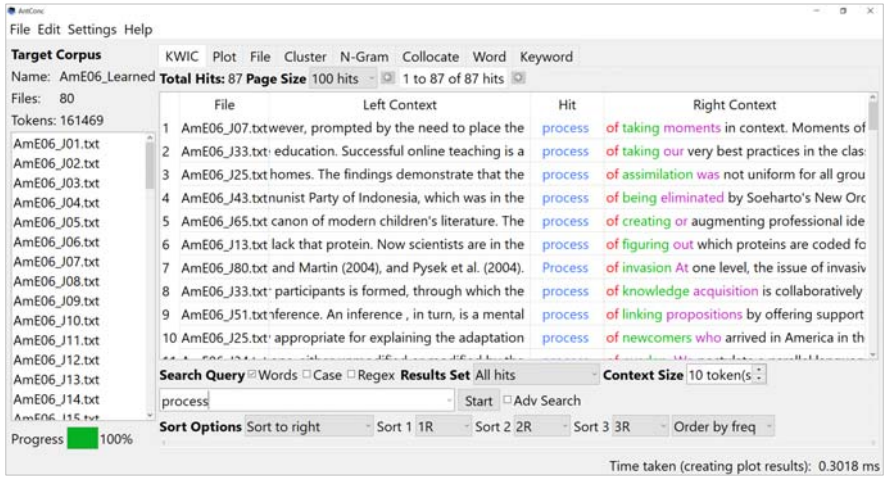
Overview of AntConc 4:

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

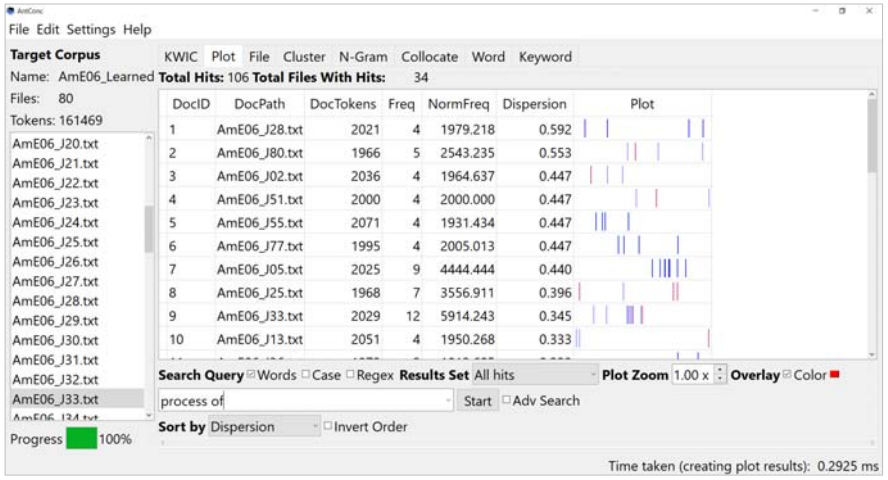
Overview of AntConc 4:

Tools and features - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword, Wordcloud



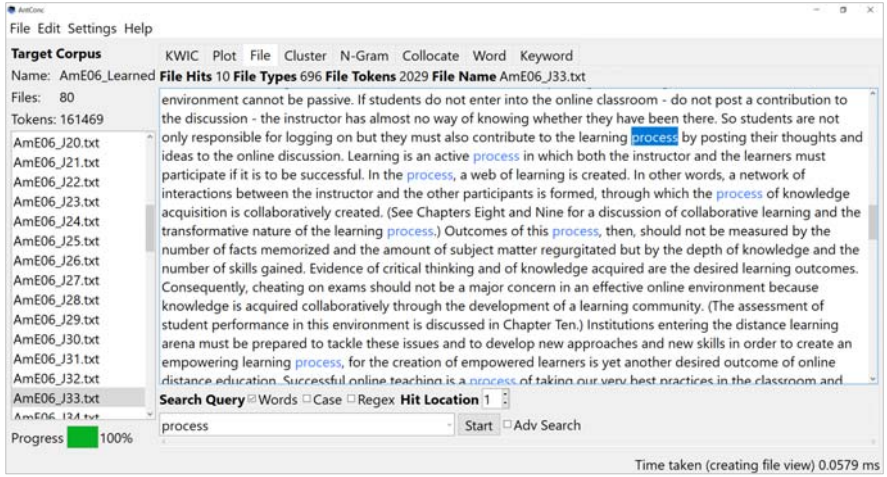
Overview of AntConc 4:

Tools and features - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword, Wordcloud



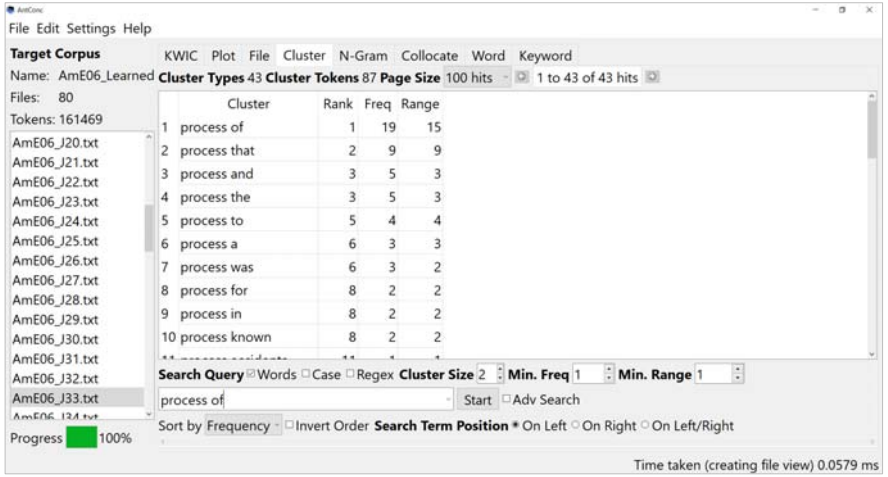
Overview of AntConc 4:

Tools and features - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword, Wordcloud



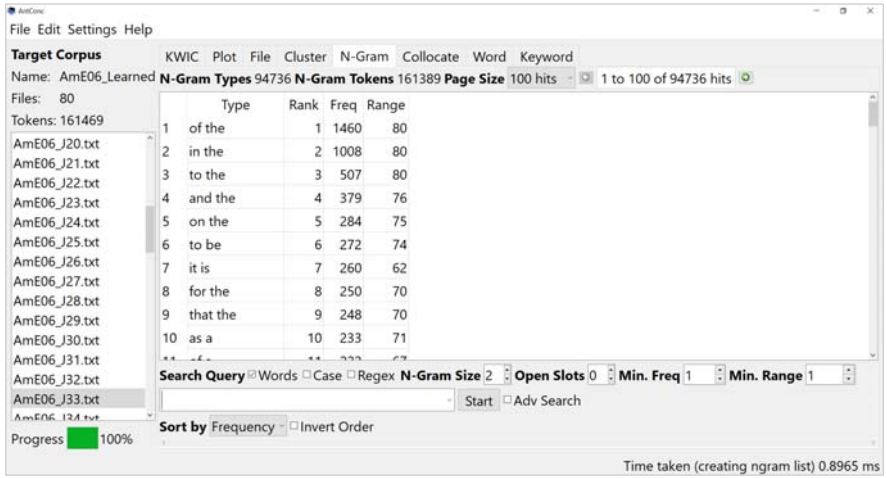
Overview of AntConc 4:

Tools and features - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword, Wordcloud



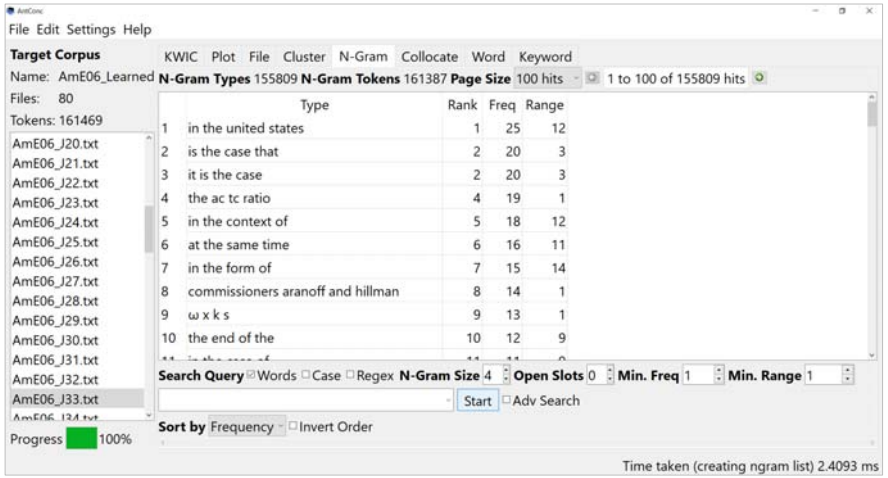
Overview of AntConc 4:

Tools and features - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword, Wordcloud



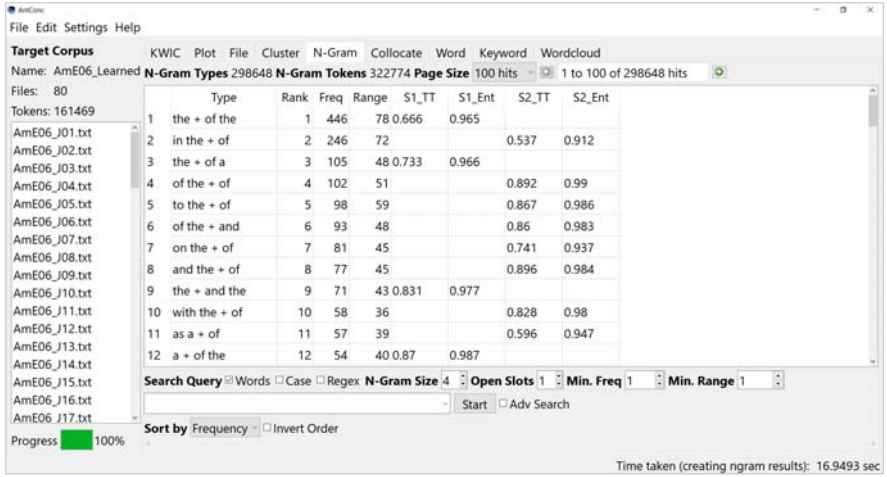
Overview of AntConc 4:

Tools and features - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword, Wordcloud



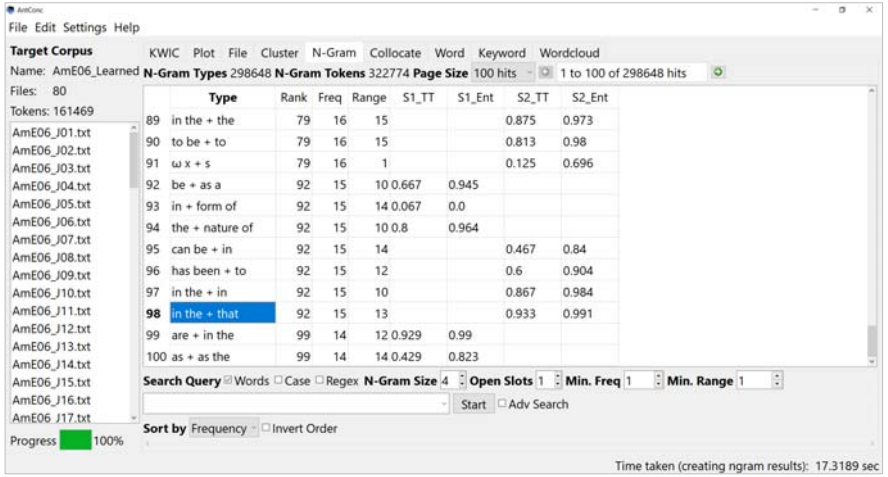
Overview of AntConc 4:

Tools and features - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword, Wordcloud



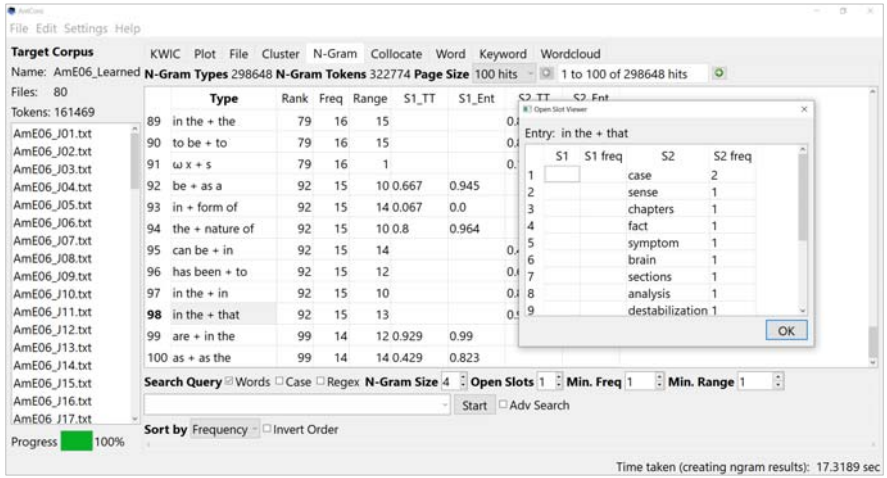
Overview of AntConc 4:

Tools and features - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword, Wordcloud



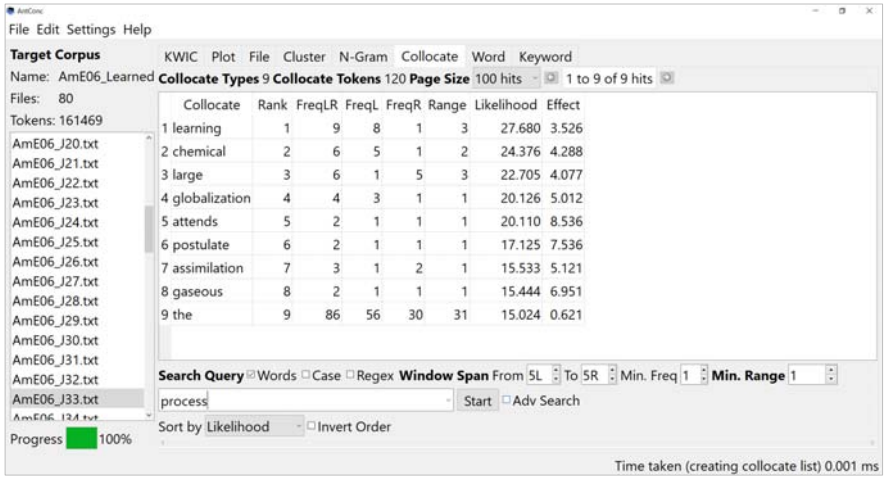
Overview of AntConc 4:

Tools and features - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword, Wordcloud



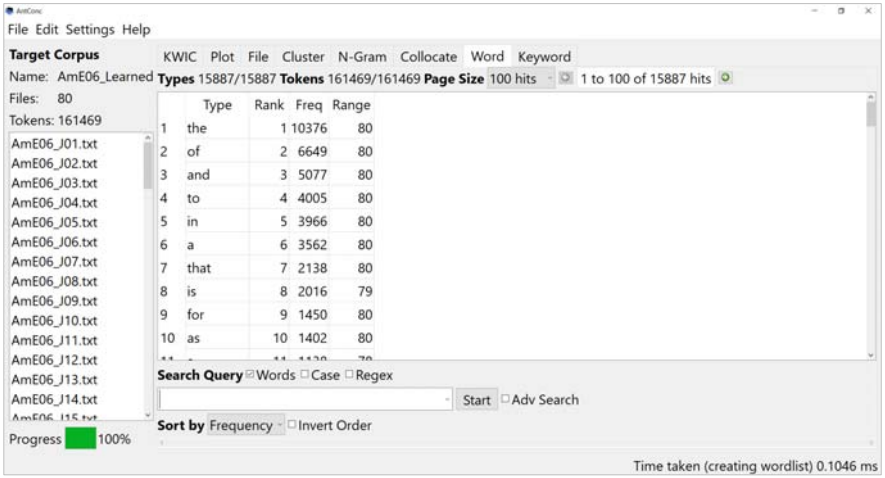
Overview of AntConc 4:

Tools and features - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword, Wordcloud



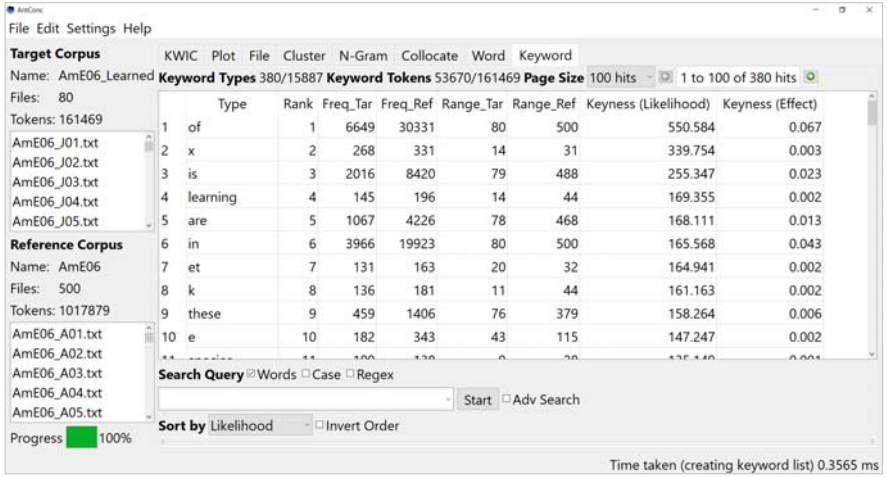
Overview of AntConc 4:

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



Overview of AntConc 4:

Tools and features - KWIC, Plot, File, Cluster, N-Gram, Collocate, Word, Keyword, Wordcloud

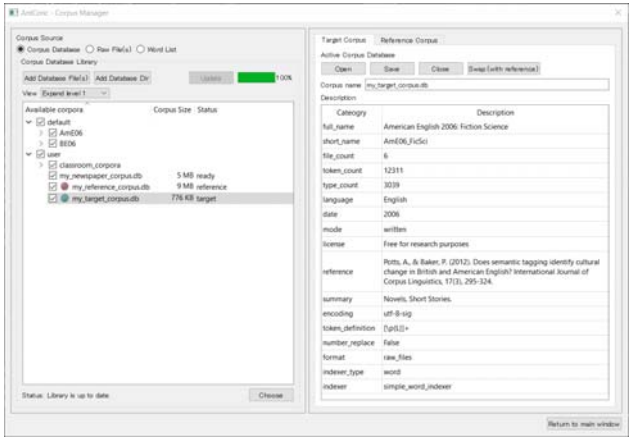
Live Demonstration

from corpora to meaningful KWIC and Keyword results

Live Demonstration:

From corpora to meaningful KWIC and keyword results

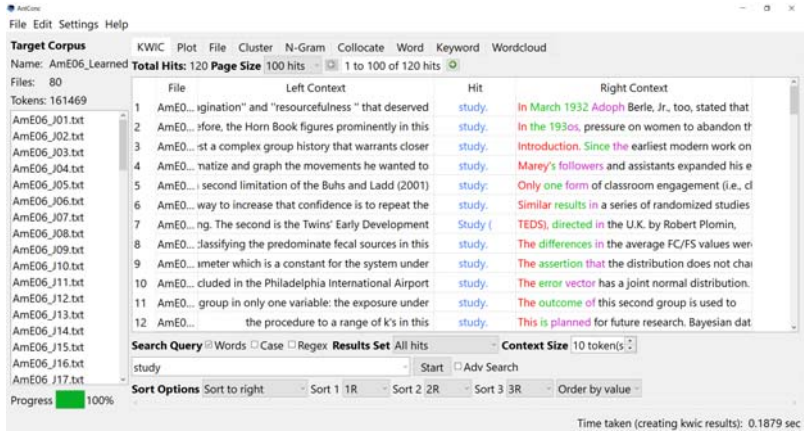
- AntConc "Corpus Manager" – online/offline repository + user corpus library



Live Demonstration:

From corpora to meaningful KWIC results

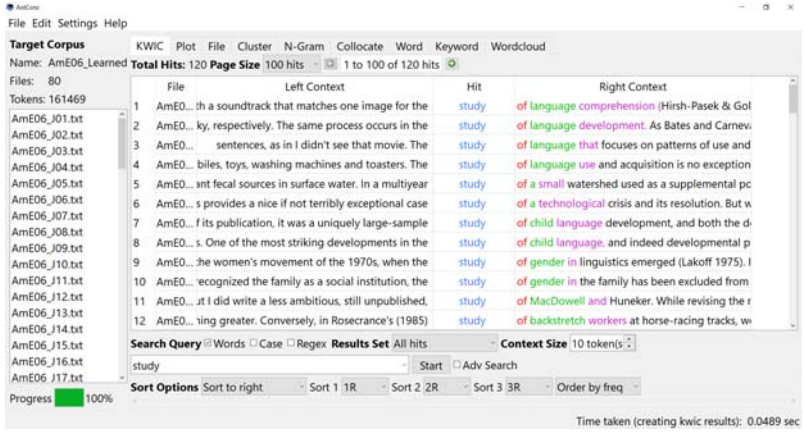
- KWIC results - value ordered (alphabetical to the right)



Live Demonstration:

From corpora to meaningful KWIC results

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



Live Demonstration:

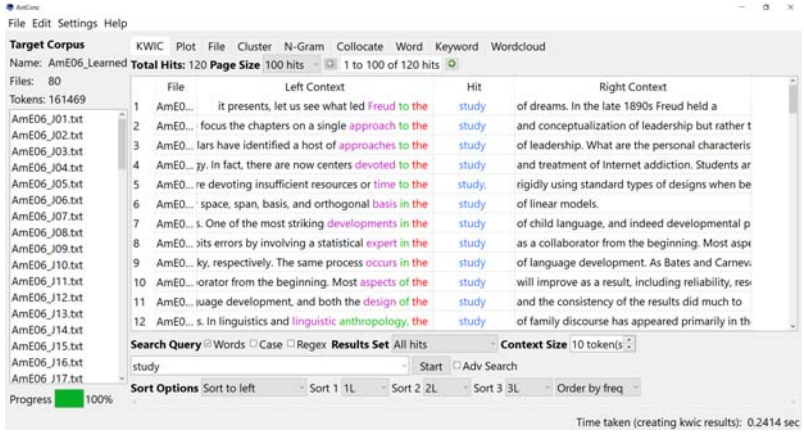
From corpora to meaningful KWIC results

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

Live Demonstration:

From corpora to meaningful KWIC results

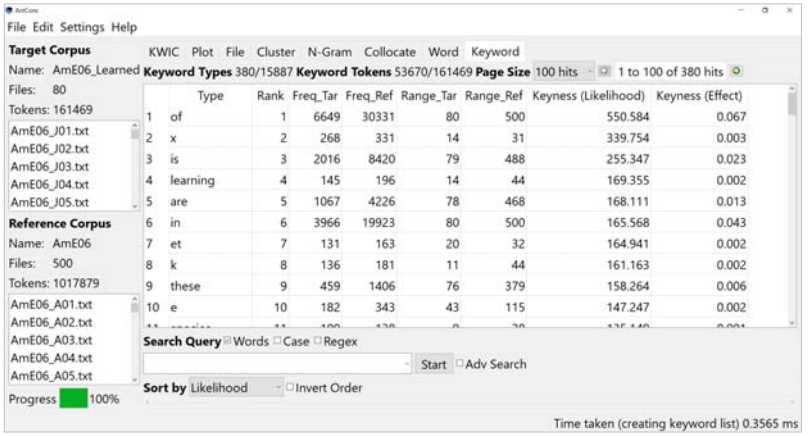
- KWIC results - frequency ordered (word patterns to the left)



Live Demonstration:

From corpora to meaningful keyword results

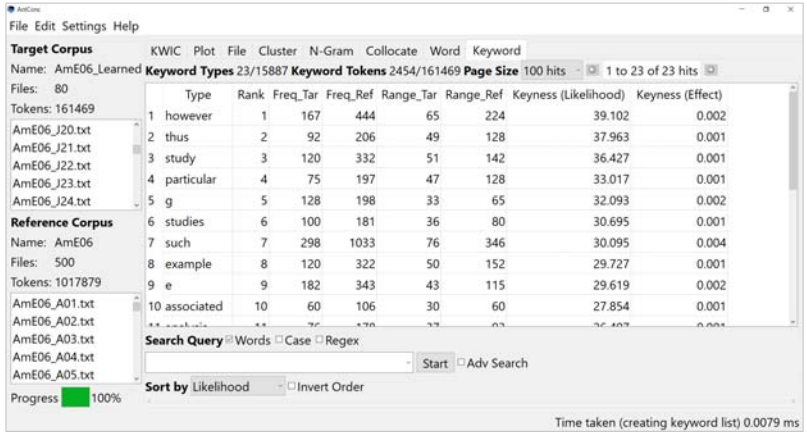
- Keyword results produced by Log-Likelihood



Live Demonstration:

From corpora to meaningful keyword results

- Keyword results produced by Text Dispersion



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

CZECH NATIONAL CORPUS

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/`

KonText

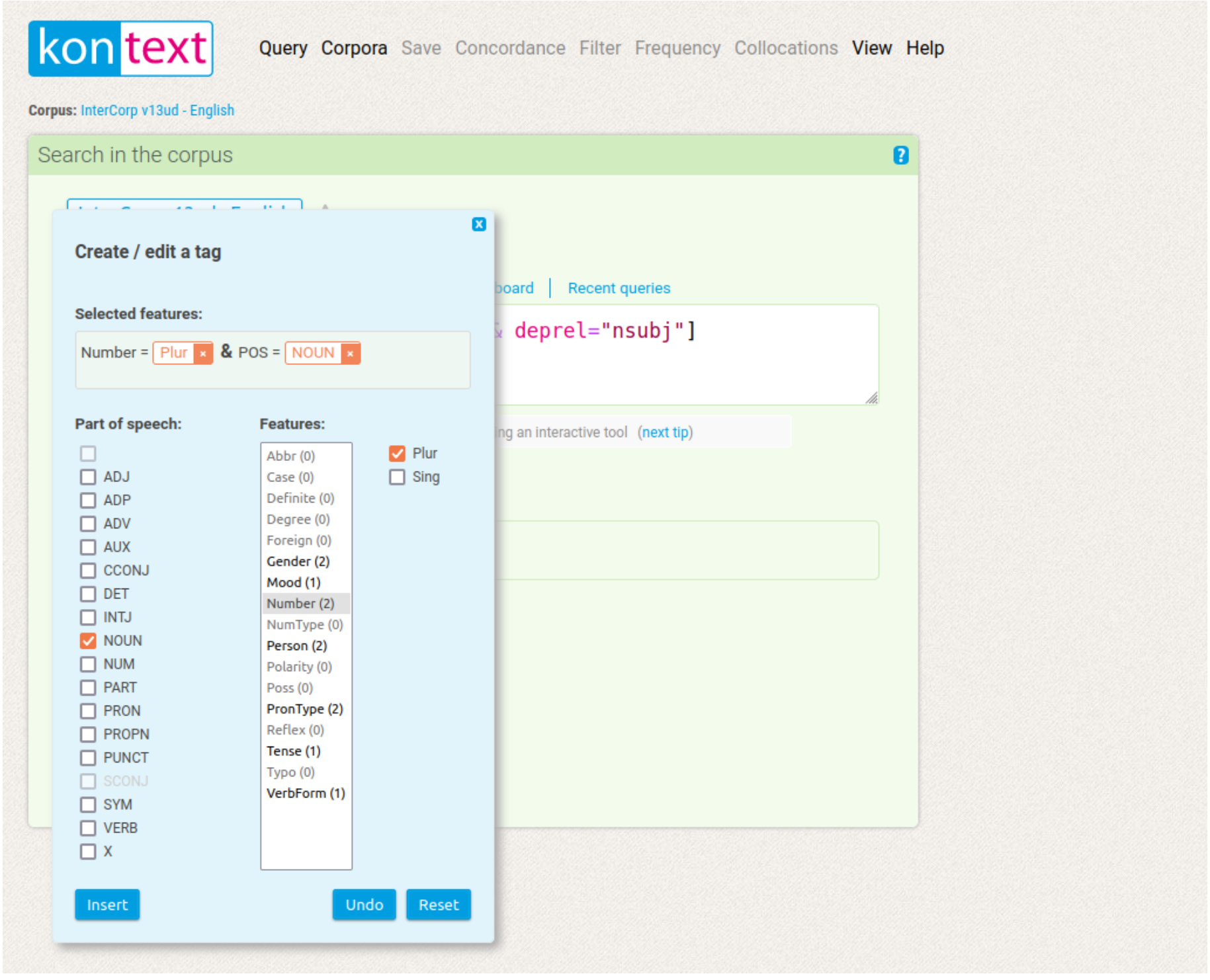
Most notable user features I

- 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
    $\Rightarrow$ easily sharable results

KonText

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)


kontext
Query Corpora Save Concordance Filter Frequency Collocations View Help
Corpus: InterCorp v13ud - English
Search in the corpus
Create / edit a tag
Selected features:
Number = Plur & POS = NOUN
Part of speech:
ADJ
ADP
ADV
AUX
CCONJ
DET
INTJ
NOUN
NUM
PART
PRON
PROPN
PUNCT
SCONJ
SYM
VERB
X
Features:
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)
Plur
Sing
Insert
Undo
Reset
Recent queries
deprel="nsubj"]
(next tip)

kontext

Query Corpora Save Concordance Filter Frequency Collocations View Help

Corpus: InterCorp v13ud - English

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

kontext

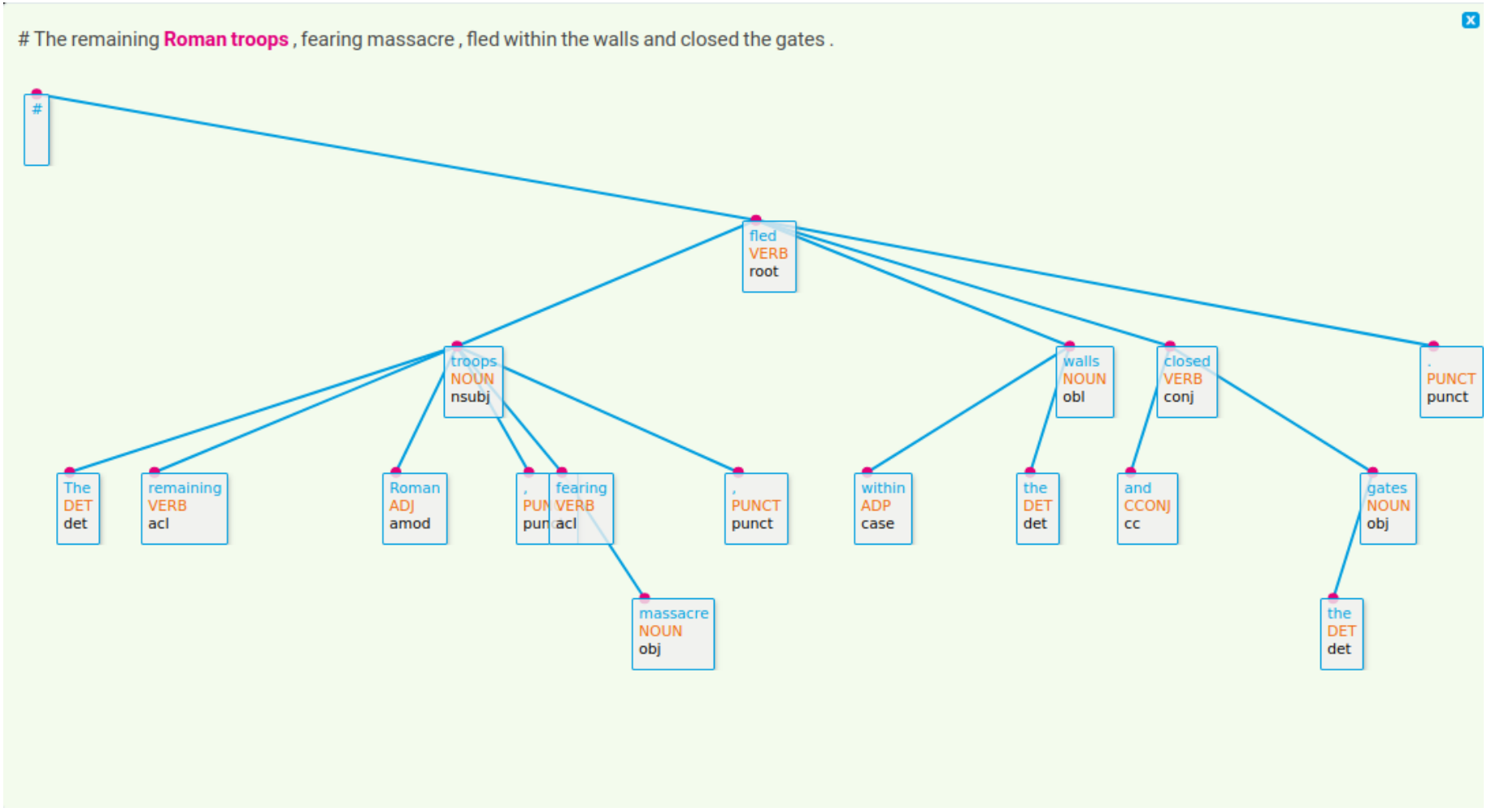
Query Corpora Save Concordance Filter Frequency Collocations View Help

Corpus: InterCorp v13ud - English | **Query**: ADJ, Number=Plur, NOUN, nsubj (321,088 hits) ▶ **Shuffle**: ✓

Hits: **321,088** | i.p.m.: **1,983.64** (related to the whole corpus) | ARF: **153,652.64** | Result is shuffled ⏮ ◀ 3 / 6,422 ▶ ⏭

Line selection: simple

Core	there is every reason to suppose that Phrygian or later	**Phoenician merchants**	reached the shores of Albion
Europarl	results from this , it is even possible that as	**local emissions**	fall , global emissions will ris
Core	battle in which the British were successful . </s><s> The remaining	**Roman troops**	, fearing massacre , fled with
Subtitles	to a war . … Police in Dallas arrested a	**young couple today … complaints**	by neighbors led them to dis
Core	. </s><s> She told herself that she had to endure a	**few hours without her nerves**	betraying her , until her granc
Subtitles	released at once to return to Scotland . </s><s> I am	**sure your families**	are missing you greatly at th
Acquis	to any party specifying such a previous year , its	**national average annual transboundary fluxes**	or national average annual e
Subtitles	… </s><s> Listen . </s><s> Thank you . </s><s> Nonsense . </s><s> What are	**good friends**	for anyway , huh ? </s><s> D
Europarl	to a halt as a result . </s><s> There are two	**other areas**	where we must make decisiv
Core	the heels of monstrous death . </s><s> But in a more	**elusive way the streets**	and houses of that vicinity b
Core	the London world , particularly in the novels of the	**twentieth century , where feelings**	of hopelessness and despon
Core	space of an hour , " there were so many	**miraculous events**	within its walls that the chro
Core	of the City . </s><s> In the early decades of the	**twentieth century new office blocks**	became a familiar presence
Acquis	been the case for thousands of years and , in	**earlier times , some activities**	– extensive land - use practi
Core	<s> I could hardly forget the lost center when all my	**open wounds**	cry out for the unity that gus
Core	stripped . </s><s> Now pigs were housed in it . </s><s> Perhaps	**other enemies**	of knowledge had been abou


The remaining Roman troops , fearing massacre , fled within the walls and closed the gates .
#
fled
VERB
root
troops
NOUN
nsubj
walls
NOUN
obl
closed
VERB
conj
.
PUNCT
punct
The
DET
det
remaining
VERB
acl
Roman
ADJ
amod
,
PUNCT
punct
fearing
VERB
acl
massacre
NOUN
obj
,
PUNCT
punct
within
ADP
case
the
DET
det
and
CCONJ
cc
gates
NOUN
obj
the
DET
det

kontext

Query Corpora Save Concordance Filter Frequency Collocations View Help

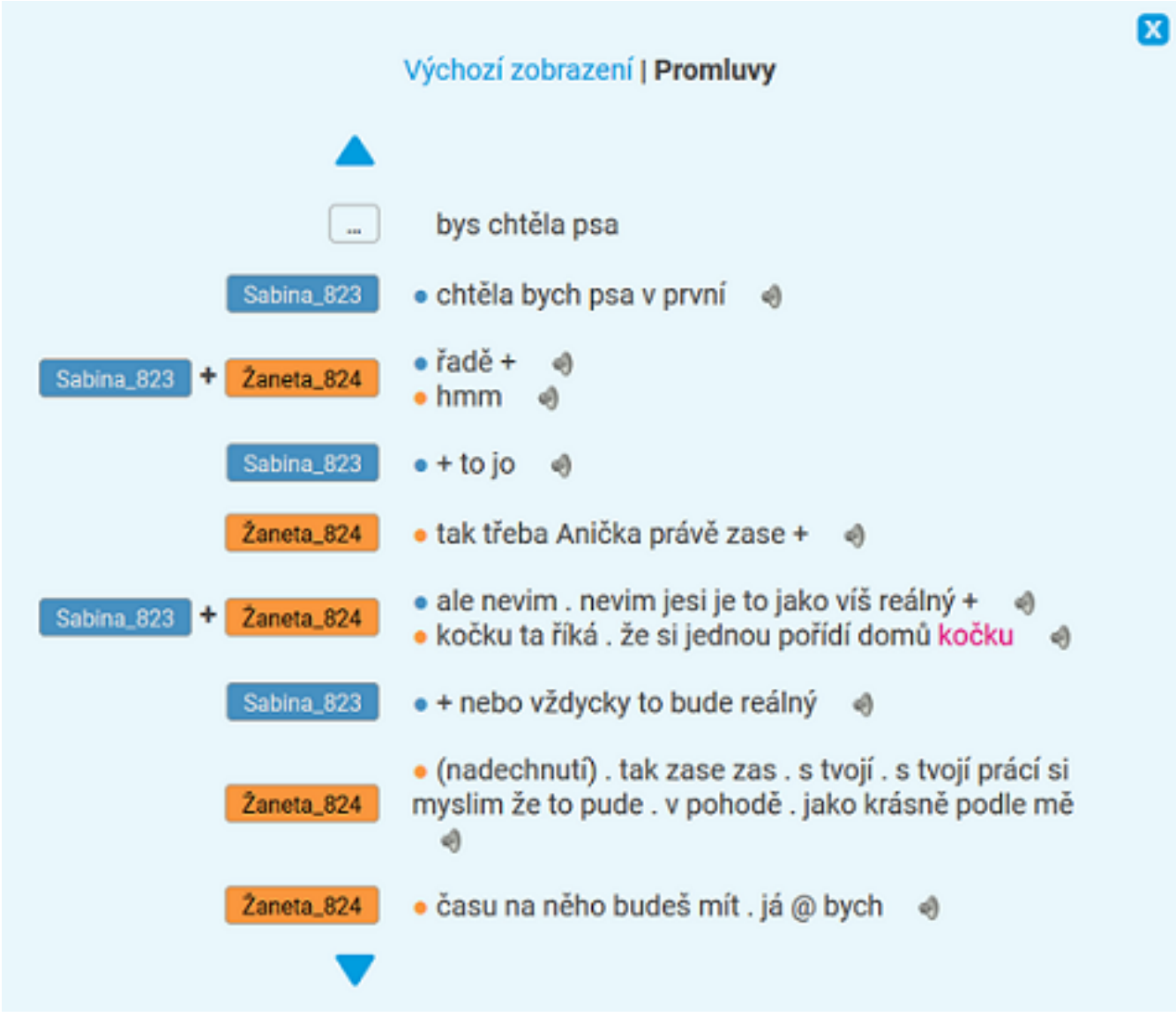
Corpus: InterCorp v13ud - English | **Query**: ADJ, Number=Plur, NOUN, nsubj (207,246 hits) ▶ **Shuffle**: ✓

Hits: **207,246** | i.p.m.: **1,280.34** (related to the whole corpus) | ARF: **86,873.94** | Result is sorted

◀◀ ◀ 2 / 4,145 ▶ ▶▶

Line selection: simple

	InterCorp v13ud - English		InterCorp v13ud - German
Acquis	<s> Appropriate **internal and external financial controls** 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 ; </s>	Acquis	<s> Geeignete interne und externe Finanzkontrollen sind im Einklang mit den international anerkannten Prüfungsnormen von der zuständigen Finanzkontrollbehörde vorzunehmen , die diese Aufgabe unabhängig erfuellen können muss ; </s>
Core	<s> But though the phalanx stayed more or less constant , the **great beasts** 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 . </s>	Core	<s> Aber obwohl die Phalanx selbst mehr oderweniger unbeweglich blieb , stürmten die großen Tiere , aus denen sie sich zusammensetzte , mit einer Geschwindigkeit von über vierzig Stundenkilometern voran , am einen Ende der Ebene aus der flirrenden Luft auftauchend und am anderen Ende ebenso abrupt wieder verschwindend . </s>
Europarl	<s> I voted for the report because I believe that these **technological advances** need to be adequately regulated in order that they do not cause systemic risk to the overall functioning of the markets . </s>	Europarl	<s> Ich habe für den Bericht gestimmt , weil ich davon überzeugt bin , dass diese technologischen Fortschritte angemessen reguliert werden müssen , damit sie kein Systemrisiko für die Gesamtfunktion der Märkte verursachen können . </s>
Core	<s> We are living in a time when flowers are trying to live on flowers , instead of growing on good rain and **black loam . </s><s> Even fireworks** , for all their prettiness , come from the chemistry of the earth . </s>	Core	<s> Wir leben in einer Zeit , wo die Blumen sich von Blumen nähren wollen , statt von gutem Regen und guter schwarzer Erde . </s>
Core	<s> Its **myriad facets** sparkled as only pricey crystal can and splashed miniature rainbows onto the walls . </s>	Core	<s> Die zahllosen Facetten funkelten , wie es nur bei wirklich edlem Kristall der Fall ist , und überzogen die Wände mit einem Schauer aus winzigen Regenbögen . </s>
Syndicate	<s> **Muslim - majority countries** spend , on average , less than 0.5 % of their GDP on research and development , compared with five times that in the advanced economies . </s>	Syndicate	<s> Länder mit muslimischen Bevölkerungsmehrheiten geben durchschnittlich weniger als 0,5% ihres BIP für Forschung und Entwicklung aus ; </s>



KonText user interface: depiction of utterances and overlaps in a dialogue.

Thank you for your attention!