# #LancsBox

## Large corpora, XML and automatic research reports



# Vaclav Brezina, William Platt et al. Lancaster University

# innovation in corpus linguistics.

# #LancsBox

@Lancaster University

### #LancsBox: License

#LancsBox is licensed under BY-NC-ND Creative commons license. #LancsBox is free for noncommercial use. The full license is available from: <u>http://creativecommons.org/licenses/by-nc-</u>nd/4.0/legalcode

When you report research carried out using #LancsBox, please cite the following:

- Brezina, V., McEnery, T. & Wattam, S. (2015). <u>Collocations in context: A new perspective on collocation networks</u>. *International Journal of Corpus Linguistics*, 20(2), 139-173.
- □ Brezina, V., & Platt, W. (2021). #LancsBox X [software package]
- Brezina, V., Weill-Tessier, P., & McEnery, A. (2021). #LancsBox v. 6.x. [software package]
- Brezina, V., Timperley, M., & McEnery, A. (2018). #LancsBox v. 4.x. [software package].



The British National Corpus 2014 is a 100-million-word corpus of modern-day British English. It was developed at Lancaster University. The spoken part of the corpus was created in collaboration with Cambridge University Press.

Brezina, V., Hawtin, A. & McEnery, T. (2021). The Written British National Corpus 2014 – Design and Comparability. *Text & Talk*.

□ Brezina, V., Love, R., & Aijmer, K. (Eds.). (2018). *Corpus Approaches to Contemporary British Speech: Sociolinguistic Studies of the Spoken BNC2014*. Abingdon: Routledge.

□ Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, *22*(3), 319-344.

## Statistical help

Brezina, V. (2018). *Statistics for corpus linguistics: A practical guide*. Cambridge: Cambridge STATISTICS IN University Press.



If you are interested in finding out details about statistical procedures used in corpus linguistics, refer to Brezina (2018); visit also Lancaster Stats Tools online at http://corpora.lancs.ac.uk/stats

 More materials (video lectures, exercises, slides etc.) are available: on the #LancsBox website: <u>http://corpora.lancs.ac.uk/lancsbox/materials.php</u>

### Downloading and running #LancsBox version 6.0

#LancsBox is a new-generation corpus analysis tool. Version 6 has been designed primarily for 64bit operating systems (Windows 64-bit, Mac and Linux) that allow the tool's best performance. #LancsBox also operates on older 32-bit systems, but its performance is somewhat limited. Version 6 of #LancsBox comes with an installer, which makes installation of #LancsBox even easier.

• Select and download: Select the version suitable for your operating system and download installer to your computer.



#### Run installer

Agree to security warnings on your machine – #LancsBox is safe to run – and follow the steps in the installer. Always install #LancsBox to a folder, where the tool has 'read and write' privileges such as the User folder or Desktop; On Windows, <u>never</u> install #LancsBox to Program Files.

Important note: System privileges

Please follow the instruction below for your specific operating system.

#### Windows 10

Windows 10 will display either of these two messages.

#### >Newer builds

"The app you are trying to install isn't a Microsoft-verified app.". If this warning message appears, click on 'Install anyway'.



#### MAC

Open "System Preferences" in the dock, click on "Security & Privacy".



Click on "Open Anyway" next to the message "LancsBox V6.0 Installer was blocked because it is not from an identified developer".



Click on "open" when the message "LancsBox V6.0 Installer.app" can't be opened because Apple cannot check it for malicious software" is displayed in a new window.

0	"LancsBox V5.1 Installer.app" can't be opened because Apple cannot check it for malicious software.
	This software needs to be updated. Contact the developer for more information.
	This item is on the disk image "lancsbox_macos_5_1_0.dmg". Firefox downloaded this disk image yesterday at 12:49 from corpora.lancs.ac.uk

## Practical tasks

**Task 1.** Many errors occur at a low-level e.g. when copying data from a spreadsheet. Can you spot six errors in the following dataset based on BE06, an approximately one-million-word corpus of written British English?

Word or expression	Frequency	Frequency per million
the	5,896	5,142.17
of	30,666	26,745.23
and	27,909	24,340.72
to	26,188	2,283.98
of the	6,887	6,006.47
and the	19,530	17,033.01
Words total	2,293,194	

Task 2. Create a Wikipedia corpus automatically in #LancsBox. #LancsBox v. 6 offers a new feature, which allows automatic download of texts from the web. In this task, we will explore this feature.

1. Choose a word or phrase that characterizes a topic of your interest (e.g. "English teaching", "ecology", "school tests", "poverty", "climate change").

My word/phrase is:

- 2. Go to *Google* or another search engine to search for articles in the selected newspaper. Include your word/phrase(s) in the search, e.g. "climate change" site: https://en.wikipedia.org
- 3. Locate your initial (seed) website that includes the topic of your choice e.g. <u>https://en.wikipedia.org/wiki/Climate\_change</u>
- 4. On 'Corpora' tab click on 'Corpus' under 'Create'

#### #LancsBox 6.0

KWIC		GraphColl	Whelk	Wor
Corpora				
Name:	Corpus 10			
Load data				
Corpus				
Word List				
Download				
Corpus				
Word List				
Create				
Corpus				

#### 5. Paste the Wikipedia url into the URL box and click on 'Create corpus'

Create Corpus		×
Corpus name	Climate	
Enter URLs (1 per line	e)	
https://en.wi	/ikipedia.org/wiki/Climate_change	
Maximum pages	100 🗢 Maximum levels 2 🗘 Allow ext	ernal links retrieval 🗌
Houristics sottings	None     Page number     O Date	
Theunsues settings		
Value matching URL	- 0	
Maximum value	0 👻	
Patterns	<u>^.*(\? &amp;)page={page}(&amp;.*\$ \$), ^.*(\? &amp;)pg={pa</u>	
	Create corpus	

6. Wait for the process to finish – by default #LancsBox is downloading and converting into text 100 websites at 2 levels of embedding – and explore the log and the folder with the data. Click 'Ok'.



Example of a lo	og:							
20210711202332	Starting request process for creating corpus "Climate"							
maxRequestPages=100								
maxQueryLevels=2								
allowExter	allowExternalToDomain=false							
useHeuris	tics=false							
HTTP_QU	ERIES_PAUSE_MIN_DURATION=500(ms)							
HTTP_QU	ERIES_PAUSE_MAX_DURATION=2000(ms)							
HTTP_QU	ERIES_PAUSE_NBER=10							
20210711202332	Starting query process for url of level 1 "https://en.wikipedia.org/wiki/Climate_change"							
20210711202334	Saving webpage textual content in corpus file: en_wikipedia_orgClimate_changeWikipedia_1.txt							
20210711202334	Extracted url: https://en.wikipedia.org/wiki/Climate_change							
20210711202334	Starting query process for url of level 2 "https://en.wikipedia.org/wiki/Environmental_impact_of_concrete"							
20210711202334	Saving webpage textual content in corpus file: en_wikipedia_orgEnvironmental_impact_of_concreteWikipedia.txt							
20210711202334	Extracted url: https://en.wikipedia.org/wiki/Environmental_impact_of_concrete							
20210711202334	Starting query process for url of level 2 "https://en.wikipedia.org/wiki/Shared_Socioeconomic_Pathways"							
20210711202334	Saving webpage textual content in corpus file: en_wikipedia_orgShared_Socioeconomic_PathwaysWikipedia.txt							
[]								
20210711202415	Number of pages for this process reached, https://en.wikipedia.org/wiki/Human_extinction hasn't been retrieved.							
20210711202415	Number of pages for this process reached, https://en.wikipedia.org/wiki/Adaptive_capacity hasn't been retrieved.							
20210711202415	Finished request process							
Summany								
Number of files creat	 ad for the cornus Climate in this request, 100							
Number of mes creat	ed for the corpus climate in this request. 100							
Pages hierarchy:								
(level 1) https://en.w	ikipedia.org/wiki/Climate_change							
(level 2) h	tps://en.wikipedia.org/wiki/Environmental_impact_of_concrete							
(level 2) h	ttps://en.wikipedia.org/wiki/Shared Socioeconomic Pathways							
(level 2) h	ttps://en.wikipedia.org/wiki/Blast_furnace#Process_engineering_and_chemistry							
(level 2) h	ttps://en.wikipedia.org/wiki/Man_of_sin							
(level 2) h	ttps://en.wikipedia.org/wiki/Nuclear winter							
, ,	I II I							

#### 7. Click on 'Import' to import and POS-tag your corpus.

#### Task 3. Save your pos-tagged corpus data.

Right-click on your corpus. In the dialogue, decide i) in which format you want to save your files and ii) what type of annotation you wish to include.

Then click on 'Save'.

Language English ~	Save corpus			×	1
Climate Language: English, 99 files	Format:	Horizontal	⊖ Vertic	al	dia_org2011_end_times_p 6899 tokens, 1926 types, 1803 lemr
● L-O-B Language: English, 15 files	Annotation:	POS	Headword Headword	Other	dia_orgAgricultural_land_ 1357 tokens, 671 types, 617 lemma
		:	Save	Format Tex	dia_org_al-Qa'im_Al_Muhai

Task 4. Explore and search your Wikipeadia corpus using #LancsBox.

1. Explore the size of your corpus and note it down:

Corpus size – tokens (running words):	
Corpus size – types (different words):	
Corpus size – lemmas (headword + POS category):	

2. How many times different words related to linguistics occur in the corpus?

Search term	Absolute frequency ( relative frequency)	Texts

Task 5. Create an automatic research report using #LancsBox Wizard

- 1. Download and import the LOB corpus.
- 2. Click on 'Wizard' to activate the Wizard tool.
- 3. Compare the corpus you created in Task with LOB. Decide on the types of analysis you want to include as well as on the search terms.
- 4. Click on 'Run'

#LancsBox 6.0			- 🗆 X
KWIC GraphColl	Whelk	Words Ngrams	Text Wizard
Corpora Wizard X			
Corpora to use Corpora to use Climate L-O-B	Tools to use  KWIC Settings  GraphColl Settings  Wreik  Words Settings  Nyrams Settings  Text Settings	Search terms veather change	Output Title Climate change Folder Jsersbrezina/Desktop/report5 Folder
All None	All None	File	Run
Wizard ready 1	1211	211+1	

#### Review the result.

Created by #Lanc	sBox Wizard			July 11, 2021 - 20:53		
Climate	e change				ated by	y #LancsBox Wizard July 11, 2021 - 20:53
	0				5 Sta	atistical analysis
1 Introdu This research rep analysis tool dev (Brezina 2018) to see the <u>Research</u> 2 Methor	ction port was automatically eloped at Lancaster Ur analyze and visualize o <u>Report Guide</u> .	produced by #LancsB liversity. It uses cuttir corpus data. For more	ox (Brezina et al., 2 Ig-edge technology a information and tip:	015, 2018, 2020), a corpus ind statistical sophistication on research report <u>writing</u>	far, dese l dispers se proce l t-test l ard to t wed a s	scriptive statistical analysis <u>was reported</u> in the sections above. This includes the analysis of frequency rsion (sections 3.1 and 3.3), keywords (section 3.2) and collocations (section 3.4). More details about cedures <u>can be found</u> in Brezina (2018). (t (108.81) = 3.43, <i>p</i> = 0.001) revealed a statistically significant difference between the corpora with the linguistic variable <i>weather</i> . This result <u>is visualised</u> below. Cohen's <i>d</i> (0.38, 95% CI [-0.17, 0.93]) small effect.
2.1 Data					data fo	or the analysis is available in <u>data\csv\statsAnalysis\weather ttest.csv</u> .
The study analyze	ed the following corpora	:				95% confidence limits
Table 1. Corpora us	ed					
Name	Language	Texts	Tokens	Additional	4 -	
Climate	English	99	621,589	Types: 62,603 Lemmas: 59,006		Т
L-O-B	English	15	1,007,677	Types: 48,349 Lemmas: 43,920	m -	
In the study, 2 c description of the	orpora were used of the corpora is available in o	he total size of 1,629, data\tsv\corpora.	266 running words	tokens) in 114 texts. A full	- 2	
2.2 Proced	ure					
#LancsBox (Brezina et al., 2015, 2018, 2020) software package <u>was.employed</u> to analyse the data. The following tools from the package <u>were used</u> . KWIC, GraphColl and Words. The KWIC tool generates a list of all instances of a search term in a corpus in the form of a concordance. The GraphColl tool identifies collocations and displays them in a table and as a collocation graph or network. The Words tool allows in-depth analysis of frequencies of types, lemmas and POS categories as well as comparison of corpora using the keywords technique. The following search terms <u>were used</u> . "weather", "climate" and "change".			lyse the data. The following rates a list of all instances of es collocations and displays th analysis of frequencies of rds technique. The following	0	I	
3 Results					7 -	
3.1 Genera	l overview: Freq	uency lists			Cli	limate L_O_B
						Corpora

Task 6. Review the query cheat sheet below.

# #LancsBox X query cheat sheet

Pasic quany		egular expre	ssions are in blue	
tea cup of tea adam 's apple	[word="apple.*"] [hw="apple"] [word="professional" pos	="J.*"]	Words starting with " Headword "apple" "professional" as an a	apple" djective
Smart search NOUN PROPER NOUN SUPERLATIVE SPLIT INFINITIVE	Complex query [word="banana"] [hw="sp [word="green blue"] [] [word="really"] []{1,3} [wo [word="big"][word="gree	olit"] ord="day"] m"]? NOUN	Banana split(s) Any word after "green "really" and "day" sep Any noun after "big"	n" or "blue" barated by 1-3 words or "big green"
XML element query <u></u> <u gender="male"></u> <text genre="fiction" subgenre=".*horror.*"></text>		Text inside Text inside Horror fict	<u></u> (utterance) e utterances with a male ion texts	lements e gender attribute
<b>"Within" query</b> [word="apple"] wi <s></s> within [word= NOUN within ( <s r<="" td=""><td>ithin <u gender="male"></u> ="yes"] n="1"/&gt; within <abstract></abstract>)</td><td>"apple" wit "yes" as a t nouns in fit (also works</td><td>thin male utterances Full sentence rst sentences of acaden s without parentheses)</td><td>nic abstracts</td></s>	ithin <u gender="male"></u> ="yes"] n="1"/> within <abstract></abstract> )	"apple" wit "yes" as a t nouns in fit (also works	thin male utterances Full sentence rst sentences of acaden s without parentheses)	nic abstracts

-

Task 7.	What is the query? Loo	at the CQL queries and	provide a short	descriptive label for e	each.
---------	------------------------	------------------------	-----------------	-------------------------	-------

CQL	Label
[word=".{3,}(tion tions ment ments ness nesses ity ities)"]	
[pos="V.D.?"]	
[pos="TO"][pos="R.*"][pos="V.*"]	
[pos="VB[^0].*"][pos="R.*"]{0,3}[pos="V.N"]	
[pos="VV."][pos="PP.*"]{0,1}[pos="RP"]	

**Task 8**. Open #LancsBox X. It will provide you with access to the British National Corpus 2014, a 100-million-word corpus of current British English. Explore the corpus and its individual subcorpora.

#LancsBo	x X			– 🗆	×	
<b>Q</b> my cat					(	
BNC2014	whole corpus	• 102M		(	X	
my cat	new subcorpus	Texts: 80/88,	171		1	
File	whole corpus		Node	Right		
ElanSo	- writing	he's who I named	my cat	after. Sadly my doesn "t		
ElanSo	elanguage	amaze myself with	my cat	photography skills. #ThisIsWhylTakePhotosOfBuildings Nop	)	
ElanRe	fiction	x These always calm	my cat	down, so I pre-emptively buy		
ElanRe	magazines	; definitely calmed	my cat	down when she was being		
ElanRe	newspapers	works well to keep	my cat	calm when I am away.		
ElanRe	official documents	t worked at all for	my cat.	I really thought the hype		
ElanRe	written-to-be-spoken	this is great stuff -	my cat	is so much calmer and		
ElanRe	point.	Feliway has never failed	my cat	yet Always works well and		
ElanRe		and quickly to help calm	my cat	when occasionally stressed. Cool cats		
ElanRe		Cat calmer It's like	my cat	is a different cat, so		
ElanRe		two male cats It works	My cat	is very anxious and had		
ElanRe		the level of anxiety of	my cat.	Sold by: Amazon EU S.a.r.L.		
ElanSo		vice versa? I can imagine	my cat	's 'What RU you doing?!		
ElanRe		as required. It works for	my cat	Excellent product Good product Used		
Sp0m2		as I love like sav	mv cat	l think I'd be		

Query	Frequency
my cat	
climate change	
brexit	
covid	
ADJECTIVE ADJECTIVE NOUN VERB	
[word="really"] [pos="J.*"]{1,3} [hw="day"]	

**Task 9**. Answer the following questions using #LancsBox X and BNC2014. Explore the 'Define new subcorpus' function.

IC2014 (whole of	corpus	-) 10	02M		
Define new	subcorpus	Name: no rest	rictions		Clear all
▼ Text					
mode	genre		subgenre	sample	academic publication
speech	academic		adventure fiction	beginning	book
writing	conversation		adventure stories & action	composite	journal
	elanguage		blog	end	
	fiction		business annual reports	middle	
magazines newspapers			chick lit; women's lit;		
			romance		
	official documents		chick literature;		
	written-to-be-	spoken	chick literature: women's		
academic discipline academic ty		academic type	publication date	spoken: nur	nber of speakers
agricultura	l and biological	editorial	2010	2	
sciences re archaeology re architecture		research ar	ticle 2011	3	
		review artic	le 2012	4	
			2012-01-01	5	
					OK Cancel

- 1. What is the relative frequency of the verb "rumble"? (Hint: verbs are words with part of speech beginning with "V").
- 2. Create a subcorpus of *whole* academic texts and find the proportion of these texts that contain split infinitives.
- 3. Modify your query to do the same search but with the whole corpus.
- 4. Compare the number of 1-word sentences "yes" to the number of 1-word sentences "no" in the speech subcorpus. Which is more common?
- 5. Find the number of 2-word sentences in the speech subcorpus.