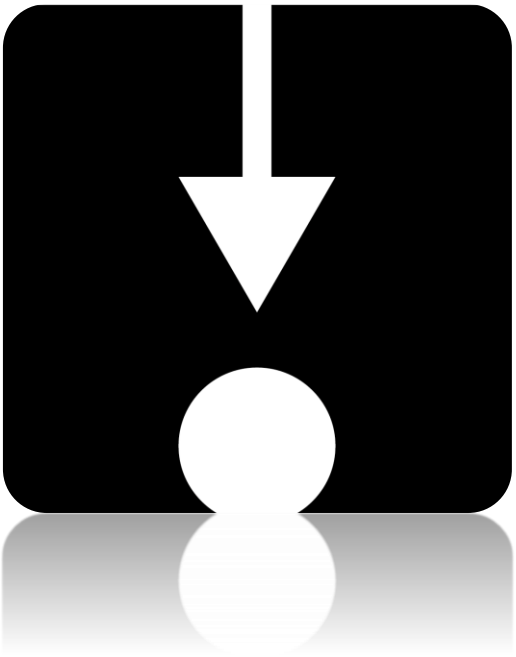
#LancsBox

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 non-commercial use. The full license is available from: http://creativecommons.org/licenses/by-nc-nd/4.0/legalcode

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

- ☐ 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.
- ☐ 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].

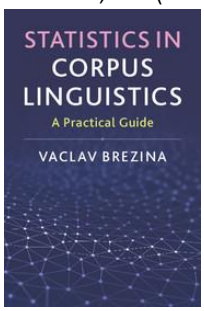
BNC2014

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 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: http://corpora.lancs.ac.uk/lancsbox/materials.php

Downloading and running #LancsBox version 6.0

#LancsBox is a new-generation corpus analysis tool. Version 6 has been designed primarily for 64-bit 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.

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

❷ 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, never install #LancsBox to Program Files.

Important note: System privileges

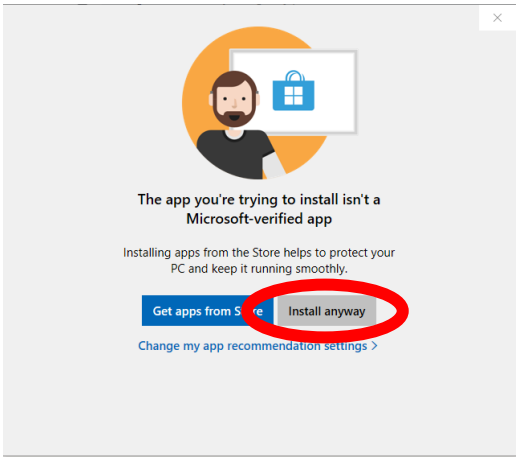
Please follow the instruction below for your specific operating system.

Windows 10

Windows 10 will display either of these two messages.

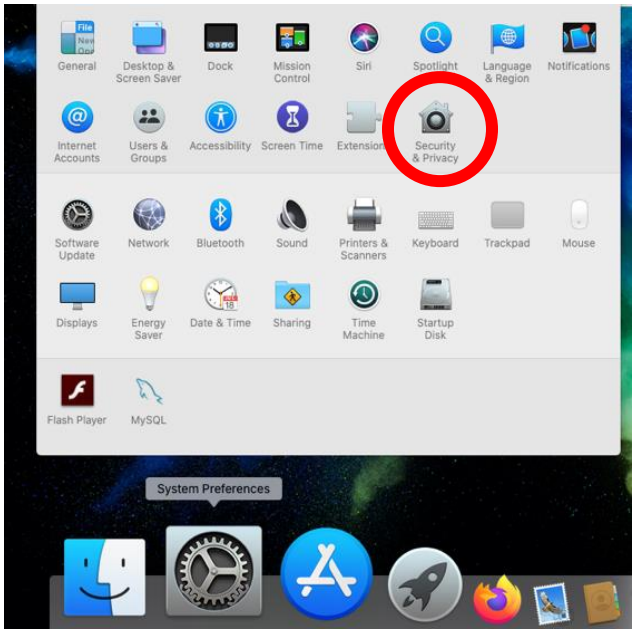
>Newer builds

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

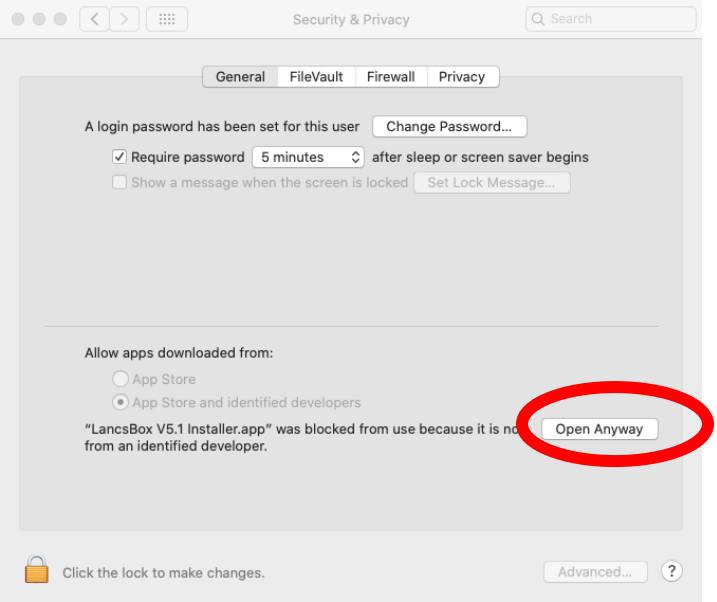


MAC

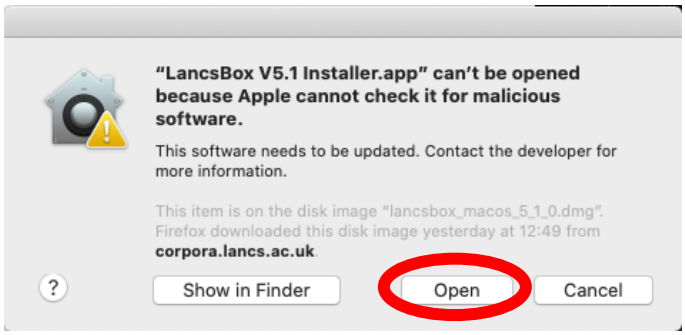
Open “System Preferences” in the dock, click on “Security & Privacy”.



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.



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.

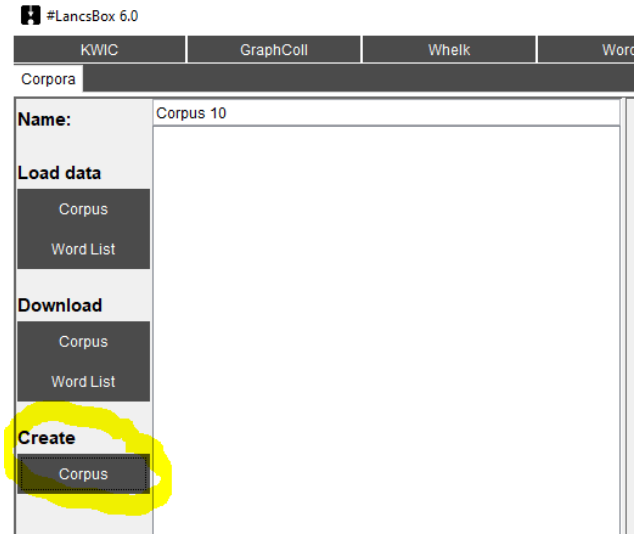
1. Choose a word or phrase that characterizes a topic of your interest (e.g. “English teaching”, “ecology”, “school tests”, “poverty”, “climate change”).

My word/phrase is: _____

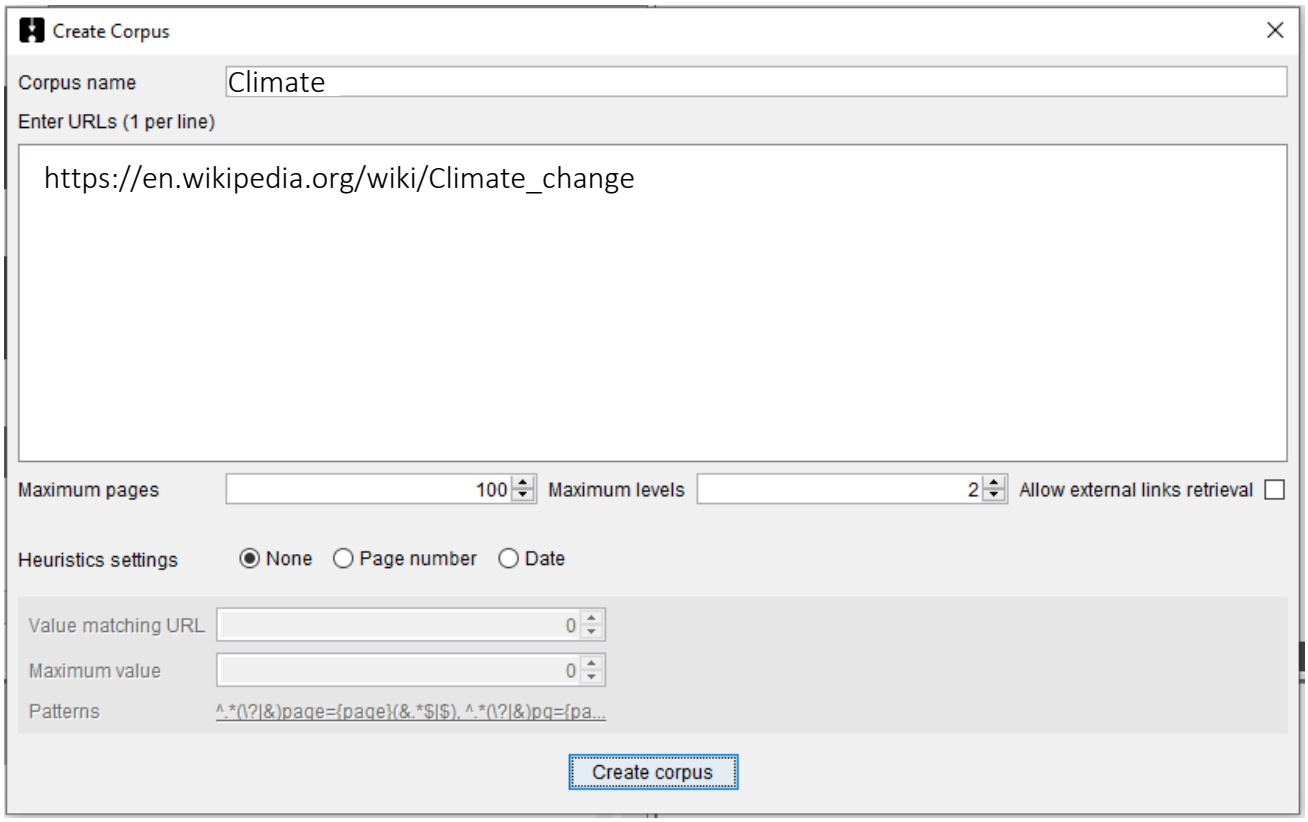
2. Go to *Google* or another search engine to search for articles in the selected newspaper. Include your word/phrase(s) in the search, e.g. "climate change" site: https://en.wikipedia.org

3. Locate your initial (seed) website that includes the topic of your choice e.g. https://en.wikipedia.org/wiki/Climate_change

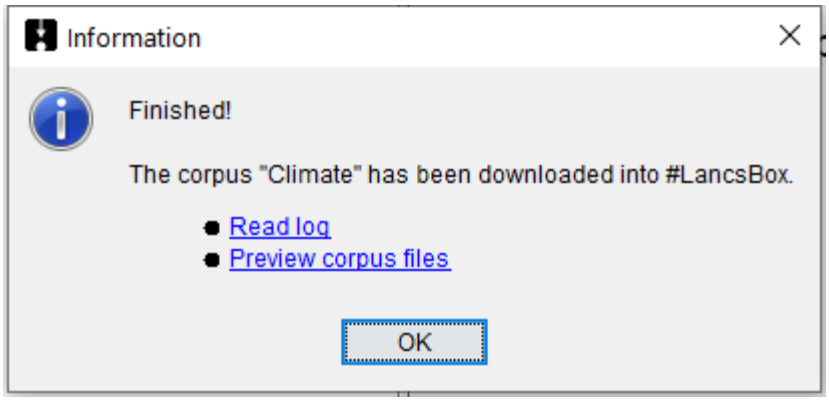
4. On ‘Corpora’ tab click on ‘Corpus’ under ‘Create’



5. Paste the Wikipedia url into the URL box and click on '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 log:

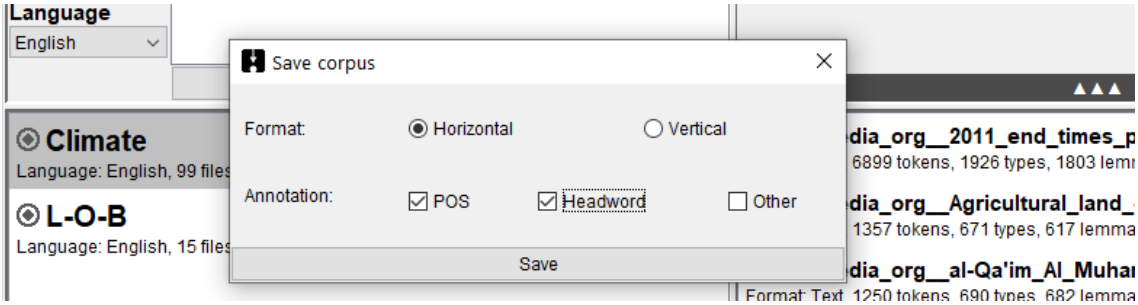
```
20210711202332      Starting request process for creating corpus "Climate"
            maxRequestPages=100
            maxQueryLevels=2
            allowExternalToDomain=false
            useHeuristics=false
            HTTP_QUERIES_PAUSE_MIN_DURATION=500(ms)
            HTTP_QUERIES_PAUSE_MAX_DURATION=2000(ms)
            HTTP_QUERIES_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_org__Climate_change_-_Wikipedia_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_org__Environmental_impact_of_concrete_-_Wikipedia.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_org__Shared_Socioeconomic_Pathways_-_Wikipedia.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


-------- Summary --------
Number of files created for the corpus Climate in this request: 100

Pages hierarchy:
(level 1) https://en.wikipedia.org/wiki/Climate_change
            (level 2) https://en.wikipedia.org/wiki/Environmental_impact_of_concrete
            (level 2) https://en.wikipedia.org/wiki/Shared_Socioeconomic_Pathways
            (level 2) https://en.wikipedia.org/wiki/Blast_furnace#Process_engineering_and_chemistry
            (level 2) https://en.wikipedia.org/wiki/Man_of_sin
            (level 2) https://en.wikipedia.org/wiki/Nuclear_winter
```

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'.



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

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

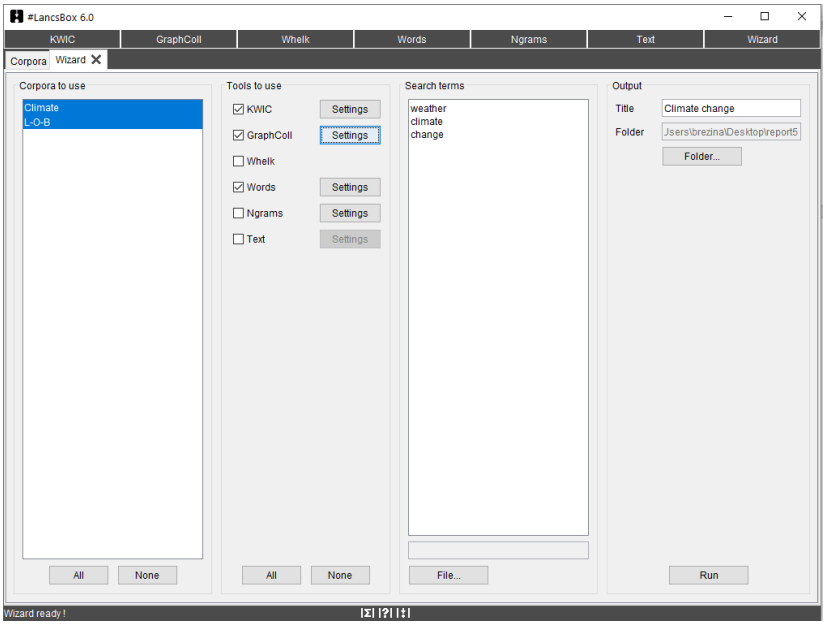
Corpus size – tokens (running words):	
Corpus size – types (different words):	
Corpus size – lemmas (headword + POS category):	

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

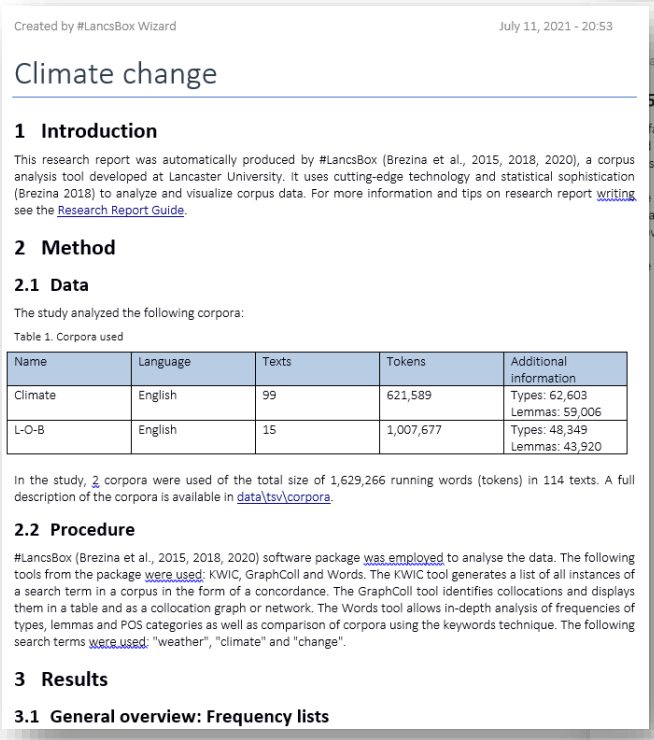
Search term	Absolute frequency (relative frequency)	Texts

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'



Review the result.



Created by #LancsBox Wizard July 11, 2021 - 20:53

Climate change

1 Introduction

This research report was automatically produced by #LancsBox (Brezina et al., 2015, 2018, 2020), a corpus analysis tool developed at Lancaster University. It uses cutting-edge technology and statistical sophistication (Brezina 2018) to analyze and visualize corpus data. For more information and tips on research report writing see the Research Report Guide.

2 Method

2.1 Data

The study analyzed the following corpora:

Table 1. Corpora used

Name	Language	Texts	Tokens	Additional information
Climate	English	99	621,589	Types: 62,603 Lemmas: 59,006
L-O-B	English	15	1,007,677	Types: 48,349 Lemmas: 43,920

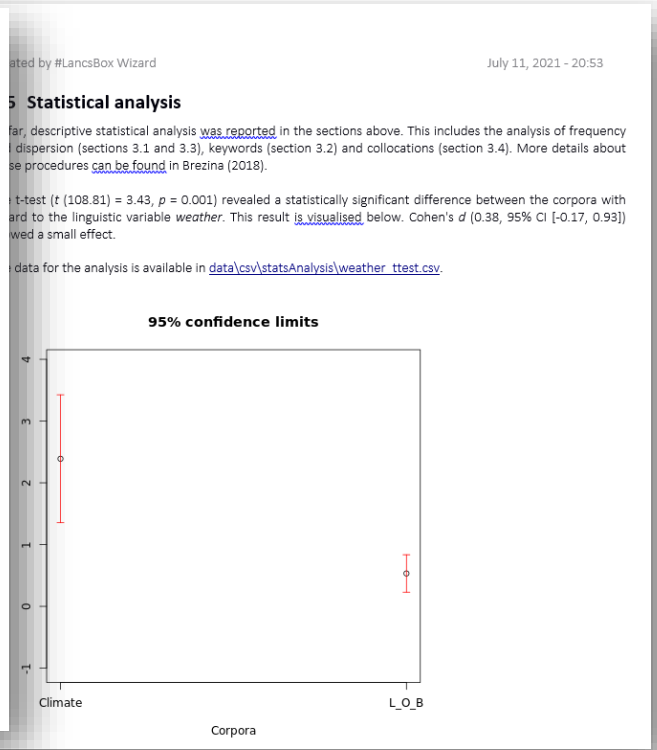
In the study, 2 corpora were used of the total size of 1,629,266 running words (tokens) in 114 texts. A full description of the corpora is available in data\tsv\corpora.

2.2 Procedure

#LancsBox (Brezina et al., 2015, 2018, 2020) software package was employed to analyse the data. The following tools from the package were used: 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 were used: "weather", "climate" and "change".

3 Results

3.1 General overview: Frequency lists



ated by #LancsBox Wizard July 11, 2021 - 20:53

5 Statistical analysis

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

t-test (*t* (108.81) = 3.43, *p* = 0.001) revealed a statistically significant difference between the corpora with ard to the linguistic variable *weather*. This result is visualised below. Cohen's *d* (0.38, 95% CI [-0.17, 0.93]) wed a small effect.

data for the analysis is available in data\csv\statsAnalysis\weather_ttest.csv.

Task 6. Review the query cheat sheet below.

#LancsBox X query cheat sheet

Regular expressions are in blue

Basic query

tea
cup of tea
adam 's apple

Smart search

NOUN
PROPER NOUN
SUPERLATIVE
SPLIT INFINITIVE

CQL

[word="apple.*"]	Words starting with "apple"
[hw="apple"]	Headword "apple"
[word="professional" pos="J.*"]	"professional" as an adjective

Complex query

[word="banana"] [hw="split"]	Banana split(s)
[word="green|blue"] []	Any word after “green” or “blue”
[word="really"] []{1,3} [word="day"]	“really” and “day” separated by 1-3 words
[word="big"][word="green"]? NOUN	Any noun after “big” or “big green”

XML element query

<u/>	Text inside <u></u> (utterance) elements
<u gender="male"/>	Text inside utterances with a male gender attribute
<text genre="fiction" subgenre=".*horror.*"/>	Horror fiction texts

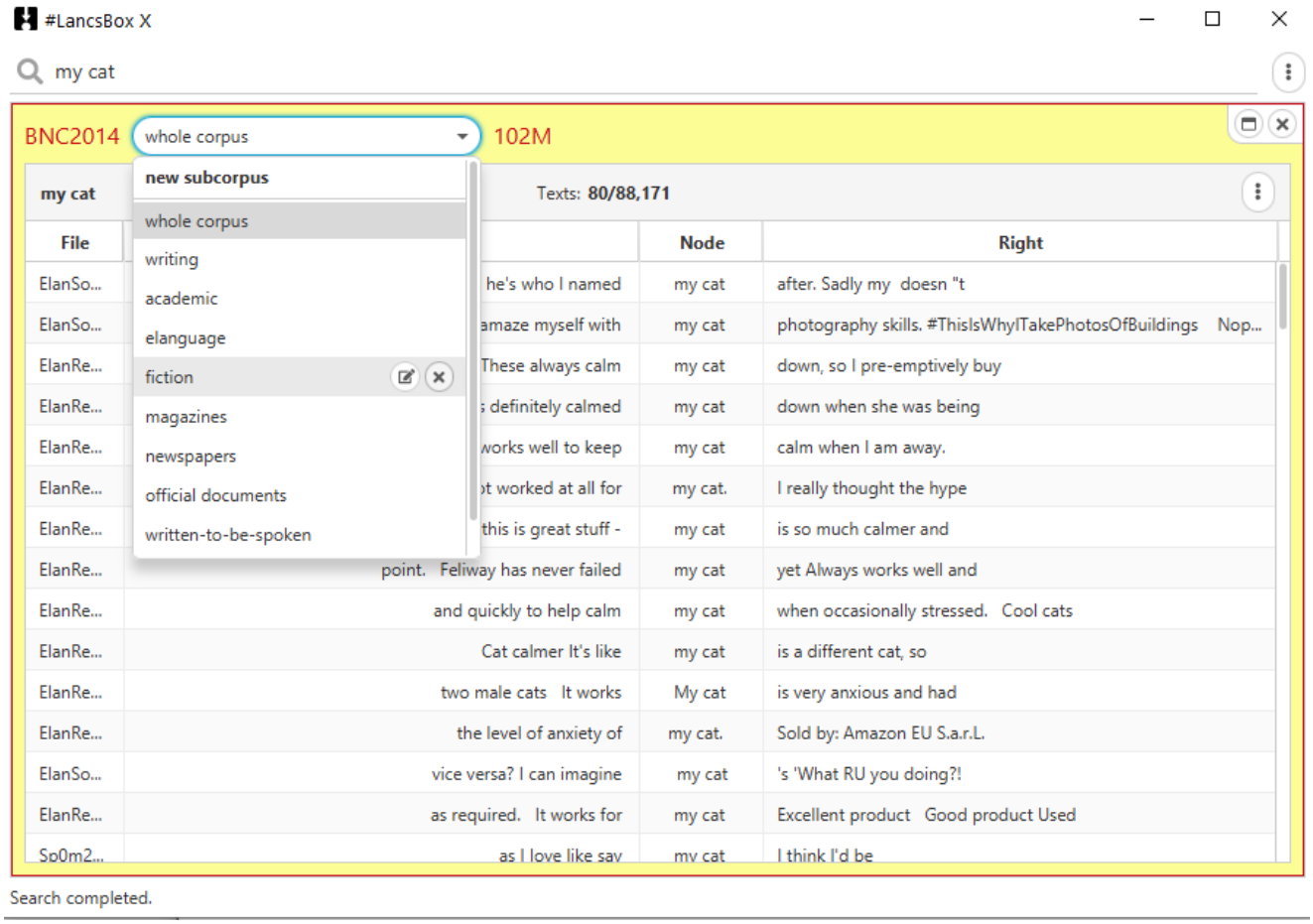
“Within” query

[word="apple"] within <u gender="male"/>	“apple” within male utterances
<s/> within [word="yes"]	“yes” as a full sentence
NOUN within (<s n="1"/> within <abstract/>)	nouns in first sentences of academic abstracts (also works without parentheses)

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

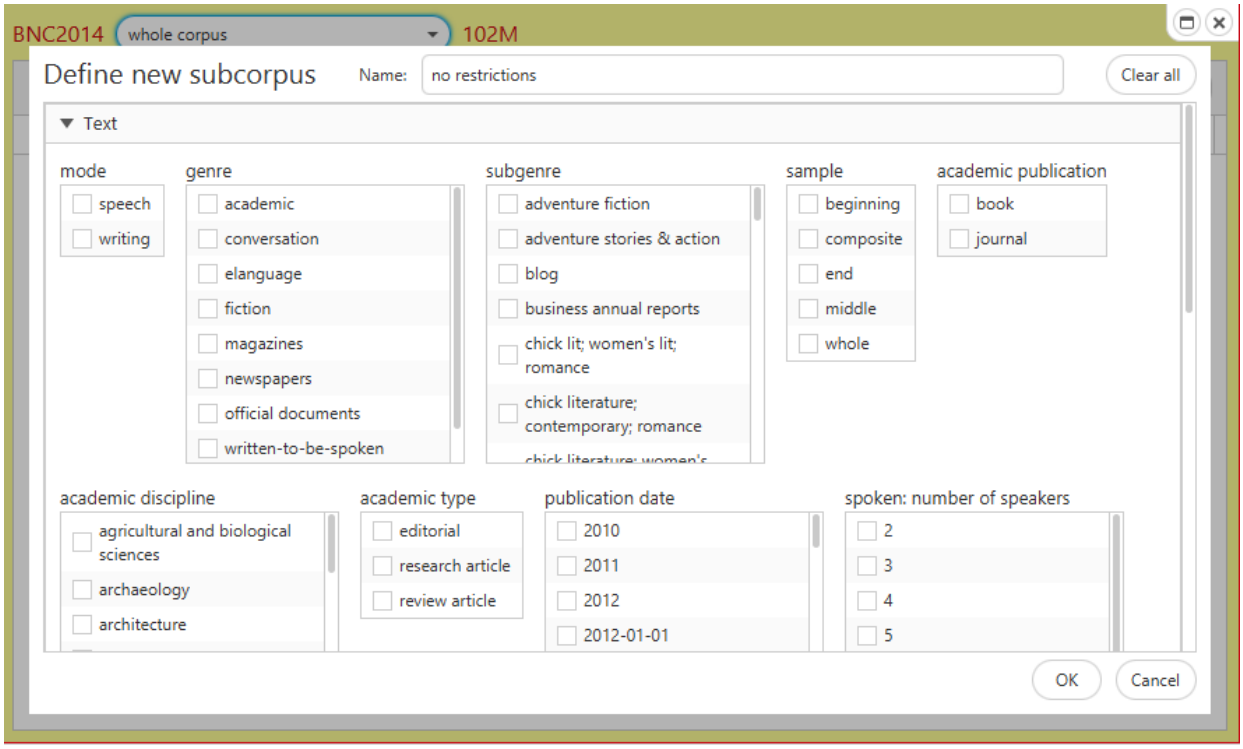
CQL	Label
[word=".{3,}(tion|tions|ment|ments|ness|nesses|ity|ities)"]	
[pos="V.D.?"]	
[pos="TO"][pos="R.*"][pos="V.*"]	
[pos="VB[^0].*"][pos="R.*"]{0,3}[pos="V.N"]	
[pos="VV."][pos="PP.*"]{0,1}[pos="RP"]	

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.



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.



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.