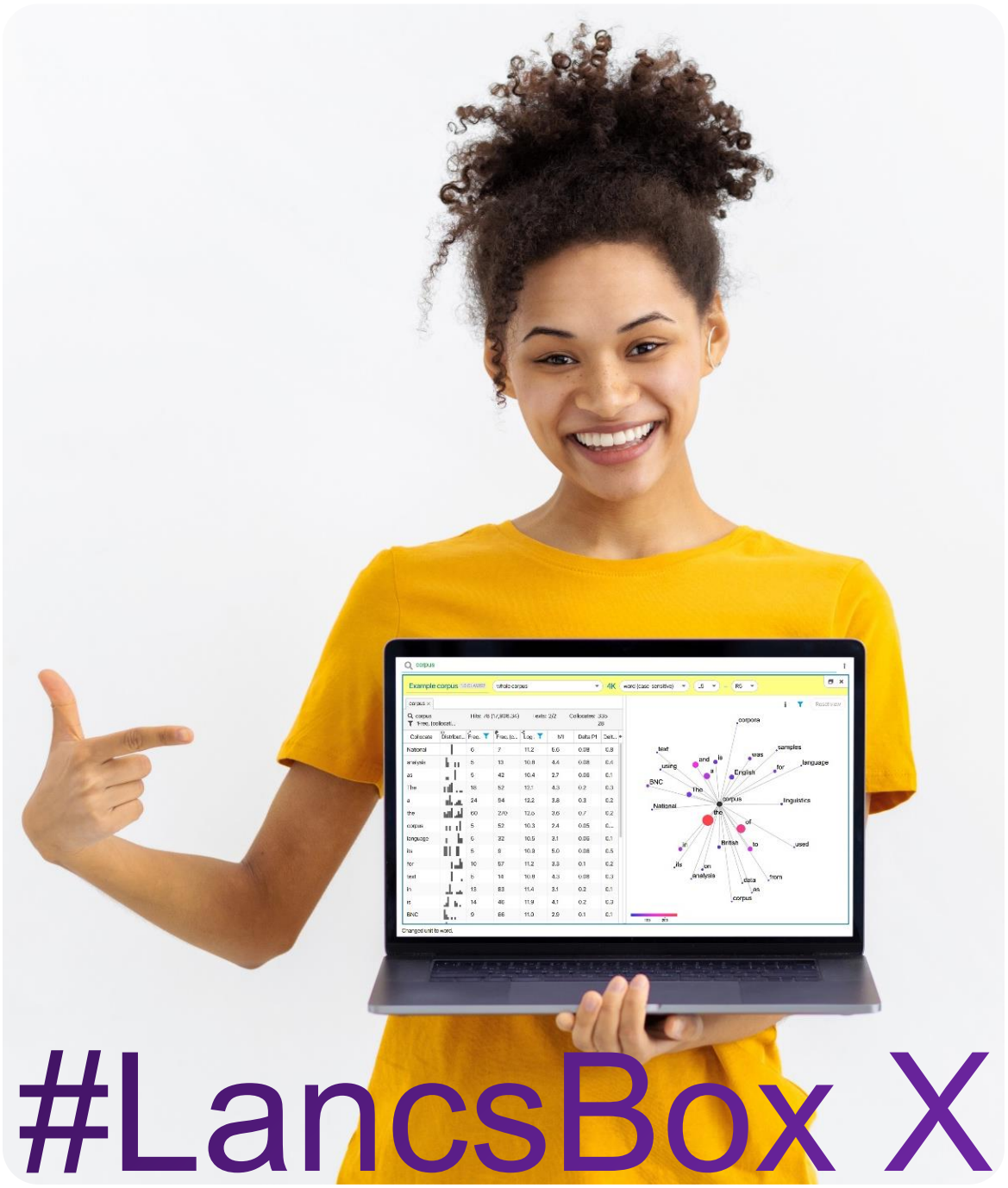


#LancsBox X

Corpus building

Exercises: building a corpus & exploring the results

In this handout, we will practice two methods of building a corpus: a manual method and an automatic method using #LancsBox X. You will then explore the data in #LancsBox X using the Text tool. The aim of this worksheet is for you to learn

- to understand the process of corpus creation
- to evaluate the quality of the data
- to understand key parameters of texts in a corpus

Text | corpus creation | Web | tokens

Task 1

Build your own newspaper mini-corpus

a) Decide on the newspaper you want to use. Choose ONE of the following British newspapers:

Newspaper	Website
The Guardian	www.theguardian.com
The Daily Mail	www.dailymail.co.uk

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

My word/phrase is: _____

c) Go to Google or another search engine to search for articles in the selected newspaper. Include your word/phrase(s) in the search,

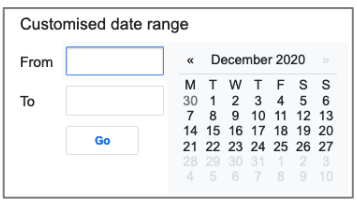
e.g. `"climate change" site:www.theguardian.com`

NB: There is no space between *site* and the web address.

d) Open the articles returned by Google (or other search engine) one-by-one and copy-paste each into a separate text document. Do this with at least 10 articles.

e) Save your text documents as plain text (.txt) in a folder.
You can use your own text editor or download the free Sublime Text. MS Word or similar word processors are not ideal but will do the job for this exercise; just remember to save the files as plain text choosing the encoding “Unicode (UTF-8)” in the saving options.

***Google* tip:** If you get many results, you can limit your search to a particular period of time by clicking on “Tools” and then on “Custom range”

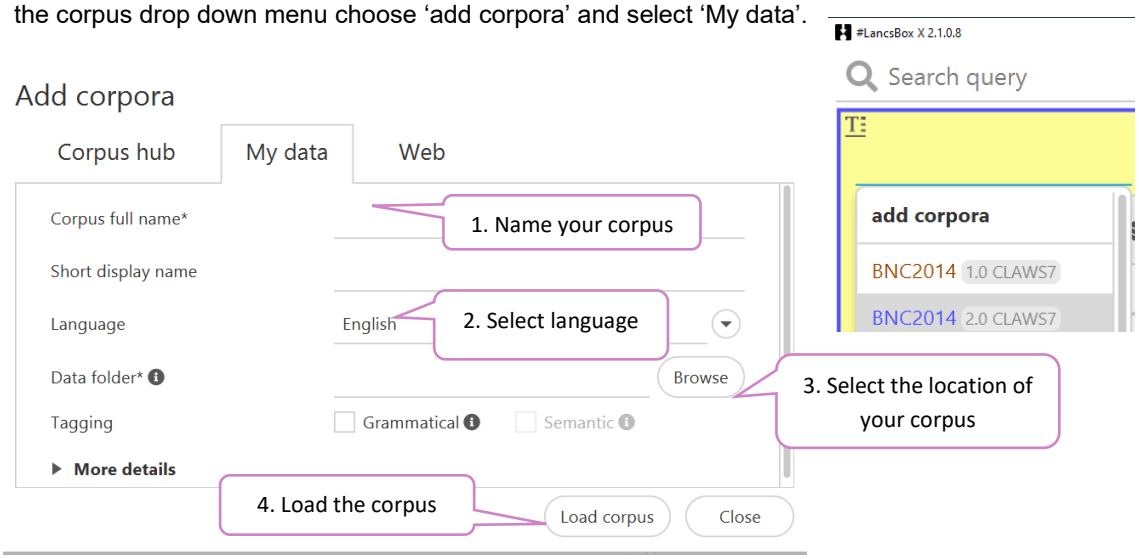


Congratulations! You’ve just created your own mini-corpus!

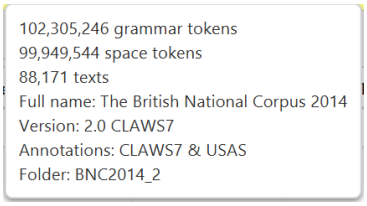
Task 2

Uploading your corpus into #LancsBox X

a) Open #LancsBox X and make sure that all your files are in a single folder e.g. MY_CORPUS. The data does not need to be in any special format – the tool accepts all major data formats (e.g. .txt, .doc., .pdf, etc.). From the corpus drop down menu choose 'add corpora' and select 'My data'.



b) Explore the corpus in the Text tool . Summary statistics are available when you hover with your mouse over the name of the corpus.



Fill in the table below with the descriptive statistics of your mini-corpus:

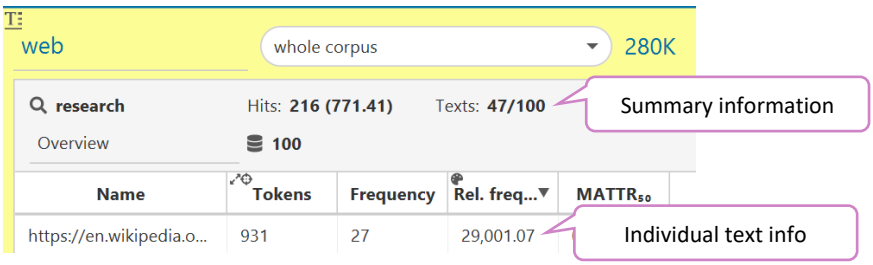
Corpus size – space tokens (strings separated by space)	
Corpus size – grammar tokens (identified by the tagger)	
Texts	

c) Tagging. When you load your own corpus into #LancsBox X, you can select 'Grammatical' and 'Semantic' tagging, which is done automatically. In the Text tool, double-click on any text and view all available annotation

Task 3

Exploring your corpus in #LancsBox X

a) Go to the Text tool in #LancsBox X and search for the frequency of your word or phrase from Task 1 in your mini-corpus.



How many times does your selected word/phrase occur in the corpus?_____

In how many texts does it appear? _____

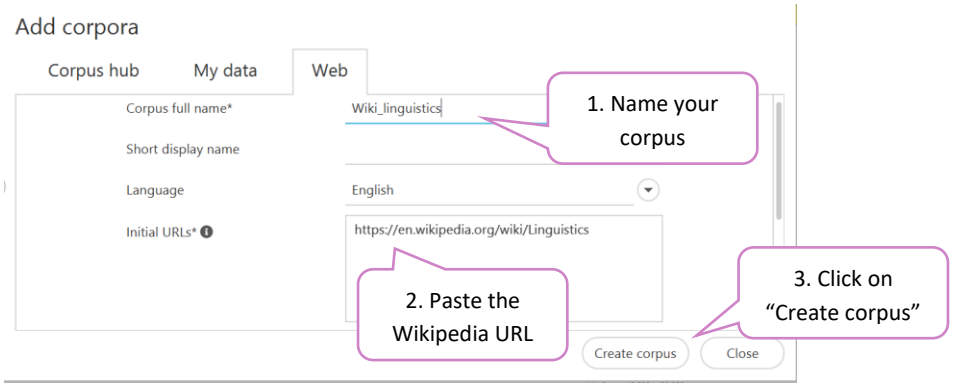
What is the text with the highest relative frequency of your word or phrase? _____

Task 4

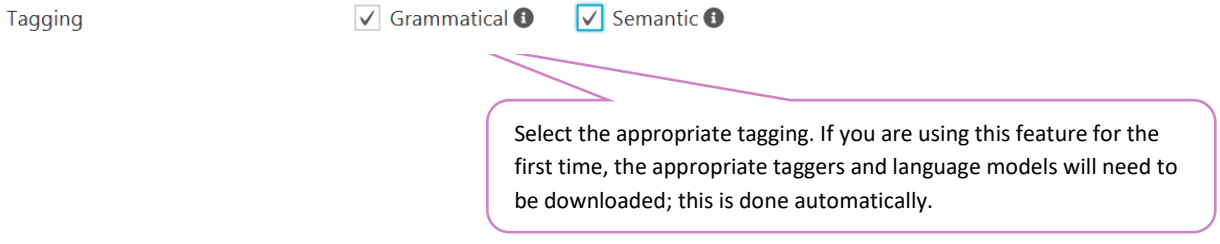
Creating a corpus automatically in #LancsBox

#LancsBox X offers a feature, which allows automatic download of texts from the web. In this task, you will explore this feature. Imagine that you want to download a corpus made of Wikipedia articles related to linguistics.

a) Locate your initial (seed) website that includes the landing page about linguistics: https://en.wikipedia.org/wiki/Linguistics From the corpus drop down menu choose 'add corpora' and select 'Web'.



b) Choose a different corpus name e.g. 'Wiki_linguistics_tagged'. Repeat the process with Grammatical and Semantic tagging on.



Task 5

Exploring your Wikipedia corpus in #LancsBox

a) Explore the size of your corpus (see instructions in Task 2b). Fill in the table below with the descriptive statistics of your Wikipedia corpus:

Corpus size – space tokens (strings separated by space)	
Corpus size – grammar tokens (identified by the tagger)	
Texts	

b) How many times do different words related to linguistics occur in the corpus?

Search term	Absolute (relative) frequency	No. of texts
`language`		
`corpus linguistics`		
…		
…		

Developed at Lancaster University