

## Building and exploring your own corpus with #LancsBox

**T** **Task 1.** Build your own newspaper mini-corpus. Follow the steps below.

1. Decide on the newspaper you want to use. Choose ONE of these:

Newspaper	Website
The Guardian	www.theguardian.com
The Telegraph	www.telegraph.co.uk
The Daily Mail	www.dailymail.co.uk
New Zealand Herald	www.nzherald.co.nz
China Daily	www.chinadaily.com.cn

2. Choose a word or phrase that characterizes a topic of your interest (e.g. "teaching English", "climate change", "rugby", "healthy lifestyle" etc.).

My topic is: \_\_\_\_\_

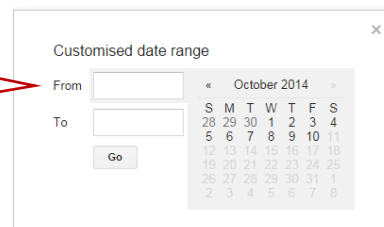
3. Go to [Google](#), [Bing](#) or [Baidu](#) to search for articles in the selected newspaper. Include your word/phrase(s) in the search,

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

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

**Google tip:**

If you get many results, you can limit your search to a particular period of time by clicking on 'Search tools'



4. 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.

5. Save your text documents as plain text (.txt).

You can use our own text editor or download a free [Notepad++](#) text editor. MS Word or similar word processors are not ideal, but will do the job ok for this exercise; just remember to save the files as plain text.

**Congratulations! You've just created your own mini-corpus!**

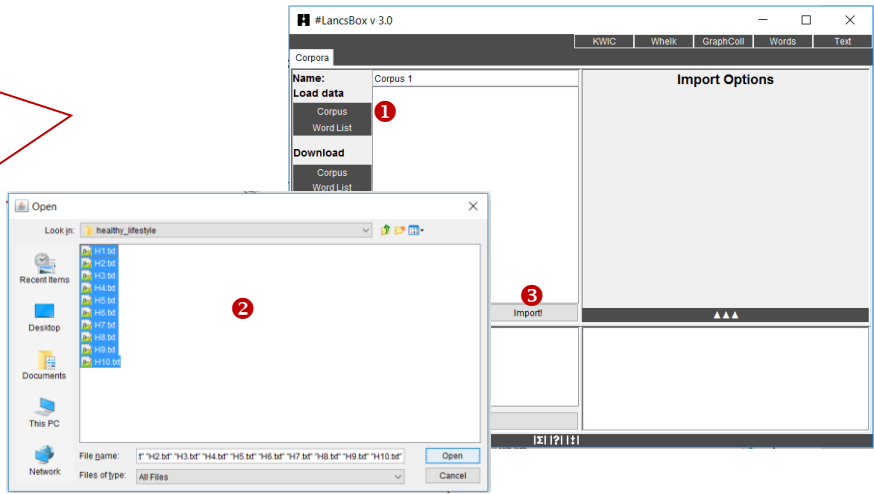
**T** **Task 2.** Use #LancsBox to analyse your mini-corpus. Follow the steps below.

1. Start #LancsBox and load your corpus files by going to the 'Corpora' tab > 'Load Corpora' > 'Corpus'. Click on the 'Import button'

A TIP: You can give your corpus a specific name (otherwise it will be just called 'Corpus 1', 'Corpus 2', etc.).

**#LancsBox tip:**

1. In the Corpora tab, left-click on 'Corpus' under 'Load data'.
2. Navigate to your corpus, select all files in the folder by holding down Ctrl (or Command) + A. Left-click 'Open'.
4. Left-click 'Import' to import your files and add POS tags.



2. 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):	

3. Search for your topic word/phrase using the Whelk tool.

- How many times does your topic word/phrase occur in the corpus? .....
- In how many texts does it appear? .....