Doing corpus linguistics with #LancsBox

In these tasks, we will become familiar with some of the functions of #LancsBox by investigating features of modern-day British English using the BNC2014 Baby+.

Tasks 1 - 2 will focus on practicing different types of **searches** using the KWIC tool.

Tasks 3 - 5 will introduce how to create and change the settings of **collocation graphs** using the GraphColl tool.

Tasks 6 - 7 will introduce how to **create a web corpus automatically** (a brand-new feature for #LancsBox v.6).

There are also **optional tasks** you can try during or after the practical session.

We will be using the **British National Corpus 2014 Baby+ edition;** this corpus is a four-million-word balanced subset of the BNC2014. Further details can be found here: http://corpora.lancs.ac.uk/lancsbox/docs/pdf/BNC2014Baby.pdf

Task 1. Searches. Go to the KWIC tool in #LancsBox and search for the following expressions in the BNC2014 Baby (provided with #LancsBox). Note down their frequencies and distributions in texts (in this case, genres).

Type of search	Search term	Occurrences (per 10k)	Number of texts
Simple	tea	**528 (1.05)**	**13**
Simple	weather	**376 (0.75)**	**13**
Phrase	mug of tea	**5 (0.010)**	**1**
Wildcard	rain*	**444 (0.88)**	**13**
Smart Search	DOWNTONER	**9,928 (19.76)**	**13**
Regex	/mate|friend/ [note that this search is case sensitive]	**1,251 (2.49)**	**13**
Regex	nick [as headword] V* [as POS]	**28 (0.06)**	**6**

Optional task: You are researching how 'hailnames' (informal forms of address) are used in British English. How might you build one query to search for *pal*, *mate* and *buddy* simultaneously? What are some issues you might face?

/pal|mate|buddy/ - mate in 'academic journal' is used as a verb – need to refine the query to ensure it is only appearing as a noun, room-mate is included, this isn't a form of address in the same way – would need to refine the query by excluding these instances.

Task 2. Applying filters. Still in the KWIC tool, search for the following expressions and apply filters. Note down their frequencies and distribution in texts.

Search term	Filter	Occurrences (per 10k)	Number of texts
tea	make [anywhere LEFT]	**19/528 (0.04)**	**9/13**
fish	chips [in R2 position]	**20/296 (0.04)**	**7/13**

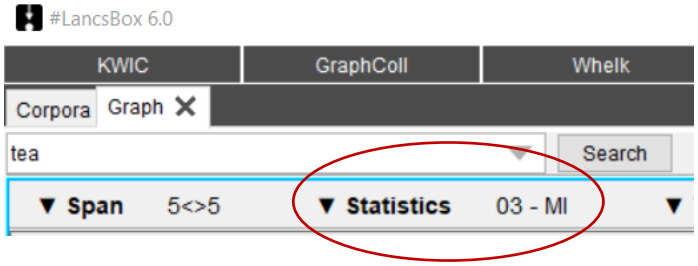
Task 3. Create a collocation graph and change settings. Go to the GraphColl tool, follow the directions and note down the frequencies and top collocates.

(a) Build a collocation graph by conducting a simple search for *tea*. What results did you get?

Frequency: 528 – Collocates: 169. Top collocate: **a**

#LancsBox v.6 manual: http://corpora.lancs.ac.uk/lancsbox/docs/pdf/LancsBox_6.0_manual.pdf

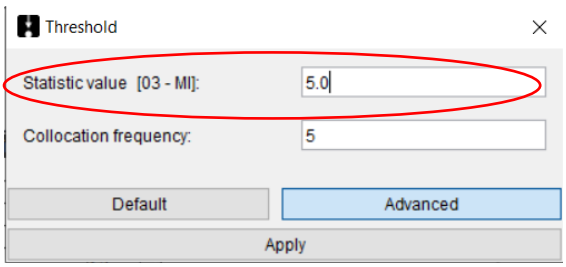
(b) Change the statistical association measure to MI score. This will delete the current graph. Search for *tea* again to create a new graph.



How has the graph changed?

Frequency: 528 – Collocates: 150. Top collocate: **fennel**

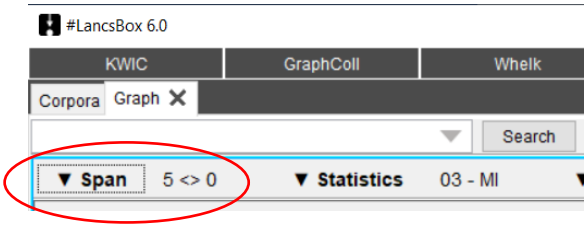
(c) Open the threshold settings and change the statistic value to 5.0 (3 and higher is considered strong for the MI score equation). Search for *tea* again.



How has the graph changed now?

Frequency: 528 – Collocates: 53 . Top collocate: **fennel**

(d) Change the window span to search for five words to the left and zero words to the right (5L, 0R). Search for *tea* again, keeping the same settings from the last steps.



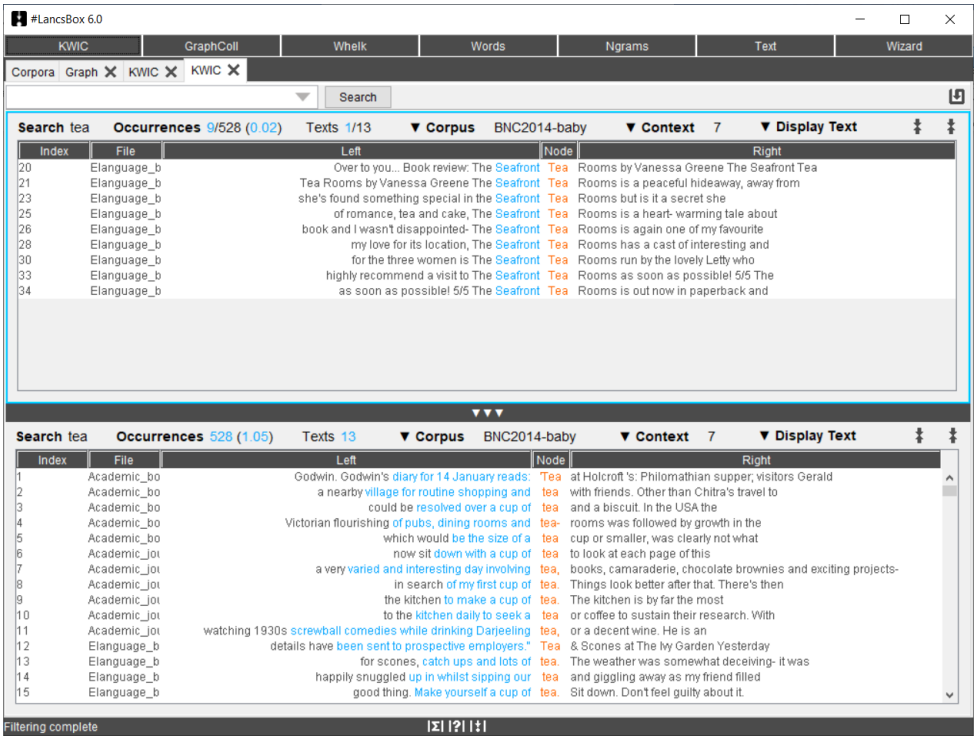
How many collocates are there now?

Frequency: 528 – Collocates: 19. Top collocate: **seafront**

Optional task: You are researching what collocations British English speakers use when talking about the weather. How might you start this study using GraphColl? What settings would you consider changing, and why? Why might you use a narrower or wider collocation window?

Using a narrower window is more appropriate for lexicogrammatical features, while a wider window can look at discourse more broadly. You'll also need to consider the size of your corpus, what search terms to use and how infrequent your search terms are as this will help you decide on the other settings, such as the thresholds.

Task 4. Combining GraphColl and KWIC view to explore collocation context. To help interpret graphs, GraphColl offers a concordance function, which displays examples of collocate use (KWIC). To display the concordance lines of a given collocation, **right click** on the collocate in the table or in the graph. These concordance lines can be expanded into the full-screen KWIC view by clicking on three arrows (▲▲▲) at the top right. In the top panel, the full-screen view displays examples of the selected collocate co-occurring with the node; in the bottom panel, all collocates are displayed.



Keeping the same settings, search for *tea* and right click on the collocate *seafront*. Explore the context in which it occurs. Comment on the association between *tea* and *seafront* in this corpus.

Looking at the concordance lines, we can see that *seafront* precedes *tea* 9 times; however, this only occurs in one genre (e-language: blog) as part of a book name even though it is a top collocate when measured by MI score. This demonstrates why we need to consider the KWIC when looking at collocates in order to make meaningful conclusions about collocations.

Task 5. Build collocation networks. GraphColl also offers the function to explore second-order collocates through building collocation networks. These are visualisations that help you explore language patterns.

(a) Search for *tea* to create a new graph (use MI score and settings: span 5L, 5R; Statistic value = 5.0; Collocation frequency = 15).

(b) Find the collocate *cup* in the graph (or table) and double click on it. This should create a collocation network similar to the one below. (n.b. you can zoom into a graph using your mouse wheel or change the text size using Ctrl +/-)

|<----------second-order collocation network--->|

|<------first-order collocation network------>|

|<----------shared collocates---------->|

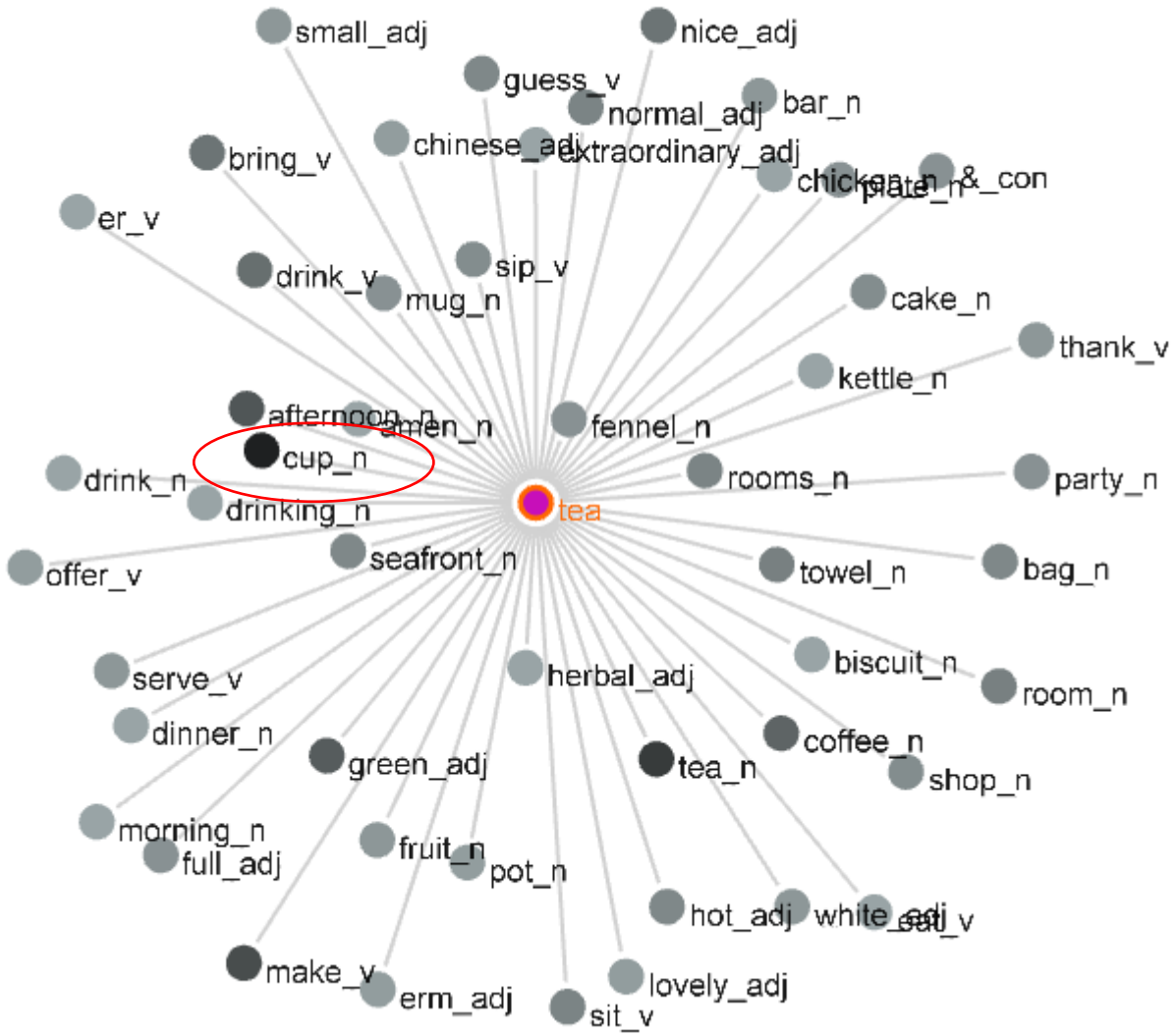
#LancsBox v.6 manual: http://corpora.lancs.ac.uk/lancsbox/docs/pdf/LancsBox_6.0_manual.pdf

(c) Comment on the connection between the collocates. Feel free to explore the network further by clicking on any of the collocates.

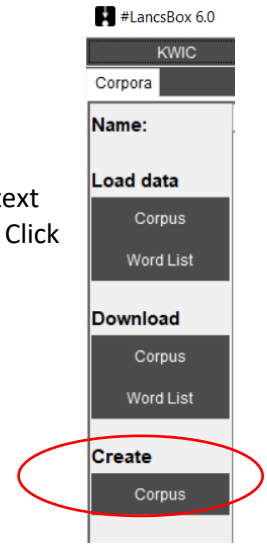
There is one shared collocate of *cup* and *tea* – this is *coffee.* From this collocation network, we can see that *cup* strongly collocates with *tea,* but it also with *rugby, world, semi* and *league* amongst other words related to sport.

Optional task: You want to investigate both collocates *cup* and *cups*. To do this, change the unit setting from type to lemma.

How does the graph change? Explore further by changing the span, statistics or threshold settings as you like.



Task 6. Create a web corpus automatically. #LancsBox v.6 offers a new feature, which allows automatic download of texts from the web.



(a) Locate your initial (seed) website that includes the landing page about your chosen topic.
(b) On 'Corpora' tab, click on 'Corpus' under create (see image)
(c) Paste the URL into the URL box and click on 'Create corpus'
(d) 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'.
(e) Click on 'Import' to import and automatically POS-tag your corpus.

Task 7. Explore and search your corpus using #LancsBox.

The following answers will depend on the website you have chosen. This example is for https://en.wikipedia.org/wiki/Language downloaded on 22.06.21

(a) Explore the size of your corpus and note it down:

Tokens (running words):	**546,540**
Types (different words):	**60,571**
Lemmas (headword + POS category):	**59,127**

(b) Explore your corpus using relevant search terms:

Search term	Occurrences (per 10k)	Number of texts
Sign	**819 (14.99)**	**51/100**
Learn	**293 (5.36)**	**100/100**
Linguistic*	**1,029 (18.83)**	**77/100**
Corpus	**29 (0.53)**	**17/100**