

Doing corpus linguistics with #LancsBox

In these tasks, we will become familiar with some of the functions of #LancsBox by investigating features of modernday 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: <u>http://corpora.lancs.ac.uk/lancsbox/docs/pdf/BNC2014Baby.pdf</u>

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 <u>collocation graph</u> by conducting a simple search for *tea*. What results did you get?
Frequency: 528 – Collocates: 169. Top collocate: a

▼ Span 5 <> 0



(b) Change the statistical association measure to MI score. This will delete the

Search

03 - MI

Statistics

current graph. Search for *tea* again to create a new graph.

#LancsBox 6.0			How has the graph changed?		
KWIC	GraphColl	Whelk	Frequency: 528 – Collocates: 150. Top collocate: fennel		
Corpora Graph 🗙					
tea		Search			
▼ Span 5<>5	▼ Statistics	03 - MI 🛛 🔻			
			(c) Open the threshold settings and change the statistic		
value to 5.0 (3 a	and higher is cons	idered strong for	the MI score equation). Search for <i>tea</i> again.		
			s the graph changed now?		
Threshold		× Freque	ncy: 528 – Collocates: 53 . Top collocate: fennel		
Statistic value [03 - MI]:	5.0				
Collocation frequency:	5				
Default	Advanced				
	Apply	(.))	and the state of the second state of the secon		
and zero words	to the right (5L, (nange the <u>window span</u> to search for five words to the left again, keeping the same settings from the last steps.		
#LancsBox 6.0		How ma	iny collocates are there now?		
KWIC	GraphColl	M/h alla	ncy: 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.



KWIC	(GraphColl	Whelk	Words		Ngrams	Text		Wizard	
rpora Grap		KWIC X								
			Search							
	-								±	
earch tea		es 9/528 (0.02)		▼ Corpus BNC2	014-bat	-		ay Text	÷	_
Index	File Elanquage b		Left Over to you	Book review: The Seafro	Node		Right	2		
	Elanguage_b			essa Greene The Seafro				ra		
	Elanguage_b			ning special in the Seafro						
	Elanguage_b			tea and cake, The <mark>Seafro</mark>						
	Elanguage_b			disappointed- The Seafro						
	Elanguage_b			or its location, The Seafro						
	Elanguage_b Elanguage b			ree women is The Seafro nend a visit to The Seafro						
	Elanguage_b Elanguage_b			possible! 5/5 The Seafro						
			400001140							
			40 0001140							
			40 0001140							
			40 0001140							
	0.000	er 528 (4.05)		***	14 babu	T Context		ay Tayt	+	
earch tea		es 528 (1.05)	Texts 13			r ▼ Context	7 ▼ Displ	ay Text	*	
earch tea	File	es 528 (1.05)	Texts 13	Corpus BNC20	Node		7 ▼ Dispi Right		*	
earch tea	File kcademic_bo	es 528 (1.05)	Texts 13 Left Godwin. Godwin's di	Corpus BNC201	Node Tea a	▼ Context	7 ▼ Dispi Right ian supper, visitors Ge		*	
earch tea	File	es 528 (1.05)	Texts 13 Left Godwin. Godwin's di a nearby village	Corpus BNC20	Node Tea a tea w	at Holcroft 's: Philomath	7 ▼ Dispi Right In supper, visitors Ge Chitra's travel to		*	
earch tea	File cademic_bo cademic_bo	es 528 (1.05)	Texts 13 Left Godwin, Godwin's di a nearby village could Victorian flourishing of	Corpus BNC201 Iary for 14 January reads for routine shopping and be resolved over a cup of tpubs, dining rooms and	Node Tea a tea w tea a tea- m	at Holcroft 's: Philomath vith friends. Other than and a biscuit. In the USA ooms was followed by	7 ▼ Displ Right In supper, visitors Ge Chitra's travel to the agrowth in the		*	
Parch tea	File ccademic_bo ccademic_bo ccademic_bo ccademic_bo ccademic_bo	es 528 (1.05)	Texts 13 Left Godwin. Godwin's di a nearby village could Victorian flourishing of whi	Corpus BNC20* iary for 14 January reads for routine shopping and be resolved over a cup of pubs, dining rooms and the vould be the size of a	Node Tea a tea w tea a tea- m tea c	at Holcroft 's: Philomath with friends. Other than and a biscuit. In the USA ooms was followed by :up or smaller, was clea	7 ▼ Dispi Right lan supper, visitors G4 Chitra's travel to the growth in the ary not what		*	
Parch tea	File cademic_bo cademic_bo cademic_bo cademic_bo cademic_bo cademic_bo	es 528 (1.05)	Texts 13 Left Godwin. Godwin's di a nearby village could Victorian flourishing of whi n	▼ Corpus BNC20° iary for 14 January reads for routine shopping and be resolved over a cup o f pubs, dining rooms and ich would be the size of a wist down with a cup o	Node Tea a tea w tea a tea- n tea c tea t	at Holcroft 's: Philomath with friends. Other than and a biscuit. In the USA ooms was followed by up or smaller, was clea o look at each page of t	7 ▼ Displ Right lan supper, visitors Ge Chitra's travel to the growth in the arly not what his	erald	*	
Parch tea	File icademic_bo icademic_bo icademic_bo icademic_bo icademic_jou icademic_jou	es 528 (1.05)	Texts 13 Left Godwin. Godwin's di a nearby village could Victorian flourishing of whi a very varied and	Corpus BNC201 lary for 14 January reads for routine shopping and be resolved over a cup o fpubs, dhing rooms and ich would be the size of a ow sit down with a cup o interesting day involving	Node Tea a tea w tea a tea- m tea c tea to tea, b	at Holcroft 's: Philomath vith friends. Other than and a biscuit. In the USA ooms was followed by :up or smaller, was clea o look at each page off books, camaraderie, ch	7 ▼ Displ Right lan supper, visitors G4 Chira's travel to the growth in the arly not what his ocolate brownies and	erald	* ts-	
Parch tea	File scademic_bo scademic_bo scademic_bo scademic_bo scademic_jou scademic_jou scademic_jou	es 528 (1.05)	Texts 13 Left Godwin. Godwin's di a nearby village could Victorian flourishing of whi n a very varied and ir	▼ Corpus BNC20° lary for 14 January reads for routine shopping and be resolved over a cup of fpubs, dining rooms and ich would be the size of a well down with a cup of interesting day involving is each of my first cup of	Node Tea a tea w tea a tea- n tea c tea to tea, b tea, T	at Holcroft 's: Philomath with friends. Other than and a biscuit. In the USA coms was followed by up or smaller, was clea o look at each page of f ooks, camaraderie, ch 'hings look better after 1	7 ▼ Displ Right Ian supper, visitors G4 Chitra's travel to the growth in the arly not what his ocolate brownies and hat. There's then	erald	ts-	
earch tea	File academic_bo academic_bo academic_bo academic_bo academic_jou academic_jou academic_jou	es 528 (1.05)	Texts 13 Left Godwin. Godwin's di a nearby village could Victorian flourishing of Victorian flourishing of m a very varied anc ir the	Corpus BNC20* iary for 14 January reads for routine shopping and be resolved over a cup o fubus, dining rooms and ch would be the size of a west down with a cup of 9 Interesting day involving is search of my first cup o is search of my first cup o	Node Tea a tea w tea a tea- n tea c tea to tea, b tea, T tea. T	at Holcroft 's: Philomath vith friends. Other than and a biscuit. In the USA ooms was followed by :up or smaller, was clea o look at each page off books, camaraderie, ch	7 ▼ Displ Right In supper, visitors Ge Chitra's travel to the growth in the arly not what his ocolate brownies and that. There's then most	erald	ts-	
Parch tea	File scademic_bo scademic_bo scademic_bo scademic_bo scademic_jou scademic_jou scademic_jou		Texts 13 Left Godwin. Godwin's di a nearby village could Victorian flourishing of m a very varied anc ir a very varied anc ir the to ti	▼ Corpus BNC20° lary for 14 January reads for routine shopping and be resolved over a cup of fpubs, dining rooms and ich would be the size of a well down with a cup of interesting day involving is each of my first cup of	Node Tea a tea w tea a tea- ri tea c tea to tea, b tea. T tea. T tea o	at Holcroft 's: Philomath with friends. Other than and a biscuit. In the USA ooms was followed by up or smaller, was clea o look at each page oft books, camaraderie, ch Things look better after The kitchen is by far the	7 V Dispi Right ian supper, visitors Ge Chitra's travel to the growth in the ariy not what his ocolate brownies and hat. There's then most research. With	erald	ts-	
Parch tea	File cademic_bo cademic_bo cademic_bo cademic_bo cademic_lou cademic_lou cademic_lou cademic_lou cademic_lou	watching 1930	Texts 13 Left Godwin. Godwins di a nearby village could Victorian flourishing oi whi n a very varied ano ir the to s crewball comedies	▼ Corpus BNC20 iary for 14 January reads for routine shopping and be resolved over a cup o fpubs, dining rooms and be resolved over a cup o interesting day involved interesting day involved interesting day involved is act of my first cup o kitchen to make a cup o kitchen to make a cup o	Node Tea a tea w tea a tea-ru tea c tea tu tea, b tea. T tea. T tea o tea, o	It Holcroft 's: Philomath with friends. Other than ind a biscuit. In the USA coms was followed by up or smaller, was clea to look at each page of th oooks, camaraderie, ch 'ings look better after 'The kitchen is by far the or coffee to sustain theil or a decent when. He is :	7 ▼ Displ Right In supper, visitors Ge Chitra's travel to A the growth in the arly not what his ocolate brownies and that. There's then most research. With an	erald	ts-	
Parch tea	File kcademic_bo kcademic_bo kcademic_bo kcademic_bo kcademic_jou kcademic_jou kcademic_jou kcademic_jou kcademic_jou	watching 1930	Texts 13 Left Godwin. Godwin's di a nearby village could Victorian flourishing of whi n a very varied and ir the to t s screwball comedies tatils have been sent to for scor	▼ Corpus BNC20* Iary for 14 January reads for routine shopping and be resolved over a cup of pubs, dining rooms and ich would be the size of a will down with a cup of will down with a cup of he kitchen to make a cup of he kitchen daily to seek a while drinking Darjeeling Darjeding	Node Tea a tea w tea a tea- m tea c tea, b tea, b tea, b tea, b tea, b tea, c tea, b tea, c tea, b tea, c tea, b tea, c tea, b tea, c tea, c t	the Holcroft 's: Philomath with friends. Other than and a biscuit. In the USA ooms was followed by up or smaller, was cleic look at each page of th nings look better after th nings look better after the rooffee to sustain theil or a decent wine. He is : 8 Scones at The by Garthe	7 ▼ Displ Right Ian supper, visitors Ge Chitra's travel to A the growth in the arly not what his ocolate brownies and hist. There's then most rresearch. With an den Yesterday what deceiving- it was	erald exciting project	ts-	

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 +/-)



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

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

(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.



T

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 <u>https://en.wikipedia.org/wiki/Language</u> 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	



