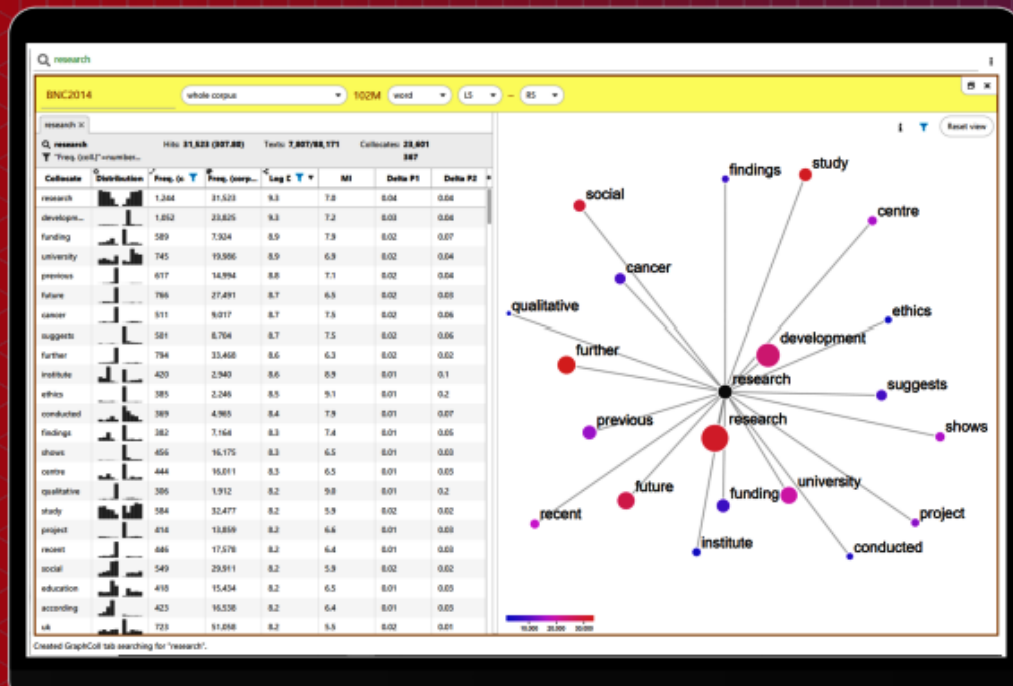


# #LancsBox Innovation in Corpus Linguistics

#LancsBox X is a powerful tool  
for the analysis of language:  
millions and billions of words.

Download it for free.



Brezina, V., Platt, W. (2024). #LancsBox X 4.0.0 [software package], [lancsbox.lancaster.ac.uk](https://lancsbox.lancaster.ac.uk)

The development of #LancsBox was supported by the Economic and Social Research Council (grant number EP/P001559/1, ES/K002155/1 and ES/R008906/1)

## Contents

1	Downloading and running #LancsBox X ....	4
2	Adding corpora .....	6
2.1	Visual summary: Corpus hub .....	6
2.2	My data .....	6
2.3	Web.....	8
2.4	Exporting corpora .....	8
3	KWIC tool (Key Word In Context) .....	9
3.1	KWIC: An overview.....	9
3.2	Multiple panels .....	10
3.3	Metadata columns .....	11
3.4	Filters .....	11
3.5	Summary table.....	12
3.6	Working with subcorpora .....	13
4	GraphColl.....	15
4.1	GraphColl: An overview .....	15
4.2	Producing a collocation graph .....	16
4.3	Reading Collocation Tables .....	16
4.4	Reading collocation graph .....	17
4.5	Extending graph to a collocation network .....	18
4.6	Shared collocates .....	19
4.7	Problems with graphs: overpopulated graphs .....	20
4.8	Reporting collocates: CPN.....	21
5	Words tool.....	22
5.1	Words: Overview .....	22
5.2	Producing frequency lists.....	23
5.3	Producing keywords and key n-grams	23
5.4	Word cloud .....	24
6	Text tool .....	26
6.1	Text: Overview .....	26
7	Searching in #LancsBox.....	28
8	spaCy POS tagset: English .....	32
9	spaCy dependency tags .....	33
10	CLAWS tagset (C7) .....	34
11	USAS semantic tagset .....	38
12	Definitions of smart searches .....	41
13	Glossary.....	44

Lancaster  
University



Developed @ Lancaster University

## #LancsBox X: License

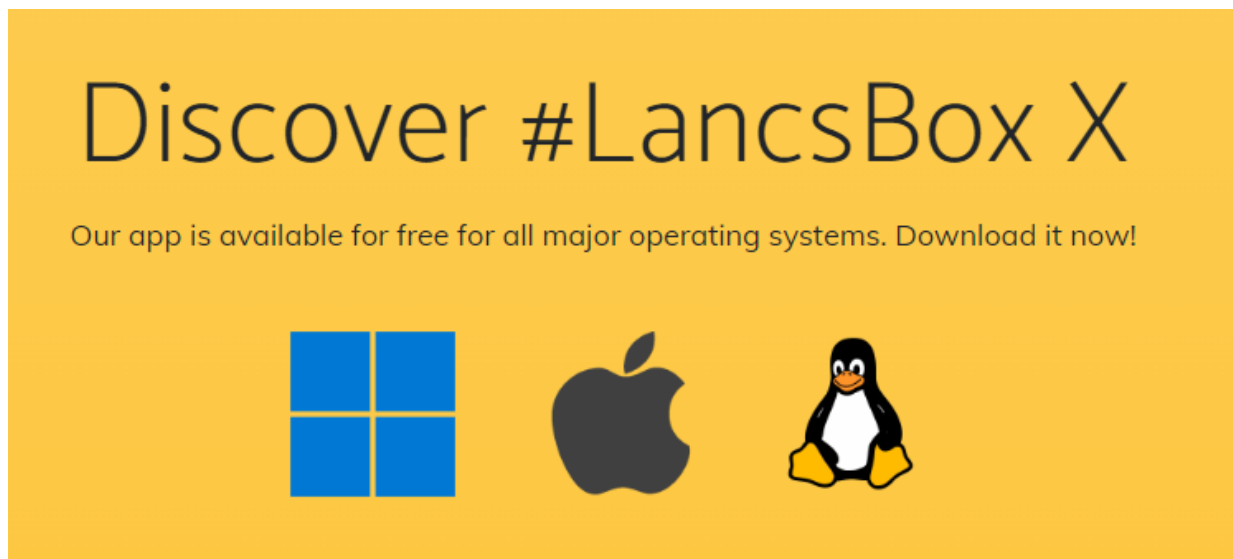
#LancsBox is licensed under BY-NC-ND Creative commons license. #LancsBox is free for non-commercial use. The full license is available from: <http://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>

### 1 Downloading and running #LancsBox X

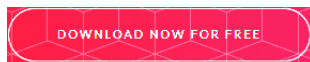
#LancsBox is a new-generation corpus analysis tool. Version X has been designed for 64-bit operating systems (Windows 64-bit, Mac and Linux) that allow the tool's best performance.

---

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



Or simply click on



#### 2 Run installer

#LancsBox is safe to run. Double-click on the installer file and follow the steps in the installer. Always install #LancsBox to a folder, where the tool has 'read and write' privileges such as the Users folder (default) or Desktop; On Windows, never install #LancsBox to Program Files.

After a typical installation, #LancsBox will be located

Windows  > This PC > Windows (C:) > Users > brezina > LancsBoxX

Mac `Macintosh HD>Users>*username*># LancsBox X`

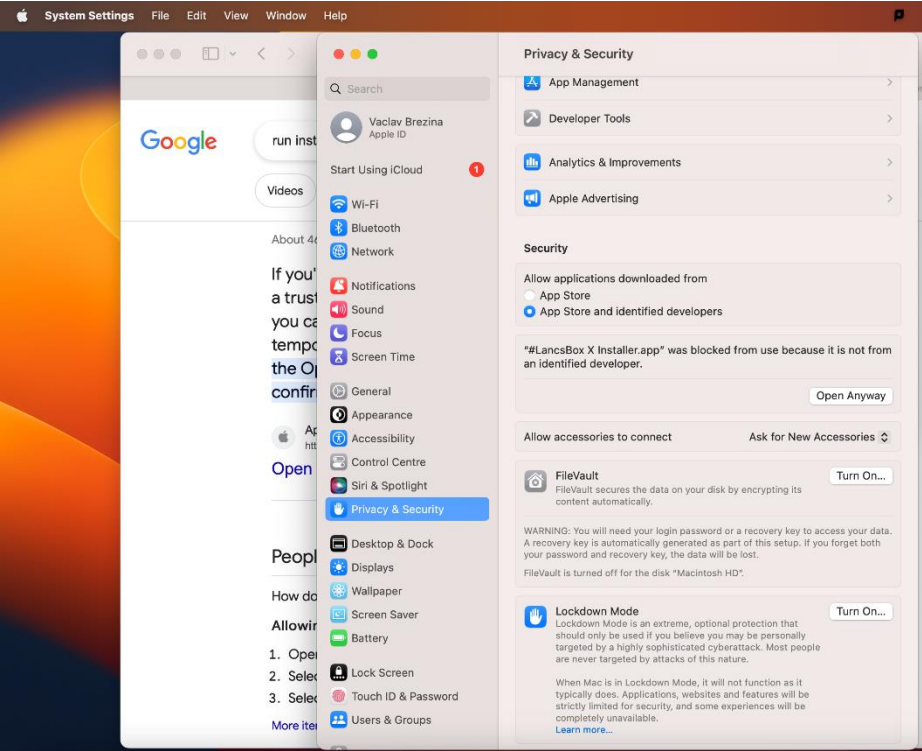
---



Please note that you may need to give the installer the privileges to run on your machine. On Windows, you might be asked for admin password.

On Mac, click on the Apple icon> System settings> Privacy & Security

Scroll down to Security, where you should be able to see ‘#LancsBox X Installer app’. Click on ‘Open Anyway’.



## 2 Adding corpora

#LancsBox X is designed for very large corpora; it natively supports XML, which allows working with rich metadata. Data can be imported into #LancsBox very easily in any format (txt, docx, pdf...). #LancsBox also has a powerful web scraping functionality.

### 2.1 Visual summary: Corpus hub

From any tool, you can add more corpora by clicking the corpus name and selecting the “add corpora” option from the dropdown menu.

**Add corpora**

Corpus hub | My data | Web

Filter: \_\_\_\_\_

Corpus name	Version	Language	Tok...
The British Nation...	1.0 CLA...	English	10...
The Lancaster Cor...	1	Chinese	83...

**You can:**

- Preview a list of available corpora in Corpus hub.
- Download existing corpora such as the BNC2014.
- Load your own data under My data.
- Create corpora from the Web.

Download corpus | Close



**Tip:** You can adjust the zoom level using the keyboard shortcuts Ctrl - and Ctrl + (Cmd - and Cmd + on a Mac).

### 2.2 My data

#LancsBox allows you to work with your own corpora. #LancsBox supports a wide range of file formats (txt, docx, pdf, pptx, xlsx...) or XML.

.txt	XML with w elements
<p>We can pick up on the last comment. Once we are in the grip of reflective thinking it is very hard, if not impossible, for us to see our ethical justifications of our ethical concepts, say, in a genuine way: we will always be drawn to the thought that this is all local. In addition, we will no longer see such judgements as embodying any sort of knowledge.</p>	<pre>&lt;?xml version="1.0" encoding="utf-8"?&gt; &lt;text id="AcaHumBk20" mode="writing" genre="academic prose"   subgenre="academic prose: humanities" subsubgenre= "academic prose: humanities: NA" publication="book" section ="NA" sample="end" source="NA" author="NA" pubDate="NA" words="6635"&gt; &lt;p n="1"&gt;&lt;s n="1"&gt;&lt;w pos="PPIS2" hw="we" class="PRON" usas= "Z8"&gt;We&lt;/w&gt; &lt;w pos="VM" hw="can" class="VERB" usas="A7"&gt;can &lt;/w&gt; &lt;w pos="VVI" hw="pick" class="VERB" usas="M2"&gt;pick&lt;/w&gt; &lt;w pos="RP" hw="up" class="ADV" usas="M2"&gt;up&lt;/w&gt; &lt;w pos= "II" hw="on" class="PREP" usas="Z5"&gt;on&lt;/w&gt; &lt;w pos="AT" hw= "the" class="ART" usas="Z5"&gt;the&lt;/w&gt; &lt;w pos="MD" hw="last" class="ADJ" usas="N4"&gt;last&lt;/w&gt; &lt;w pos="NN1" hw="comment" class="SUBST" usas="Q2:1"&gt;comment&lt;/w&gt;&lt;c&gt;.&lt;/c&gt;&lt;/s&gt; &lt;s n="2" &gt;&lt;w pos="CS" hw="once" class="CONJ" usas="Z5"&gt;Once&lt;/w&gt; &lt;w pos="PPIS2" hw="we" class="PRON" usas="Z8"&gt;we&lt;/w&gt; &lt;w pos= "VBR" hw="be" class="VERB" usas="A3"&gt;are&lt;/w&gt; &lt;w pos="II" hw ="in" class="PREP" usas="Z5"&gt;in&lt;/w&gt; &lt;w pos="AT" hw="the" class="ART" usas="Z5"&gt;the&lt;/w&gt; &lt;w pos="NN1" hw="grip" class= "SUBST" usas="A1:1:1"&gt;grip&lt;/w&gt; &lt;w pos="IO" hw="of" class= "PREP" usas="Z5"&gt;of&lt;/w&gt; &lt;w pos="JJ" hw="reflective" class= "ADJ" usas="X2:1"&gt;reflective&lt;/w&gt; &lt;w pos="NN1" hw="thinking" class="SUBST" usas="X2:1"&gt;thinking&lt;/w&gt; &lt;w pos="PPH1" hw= "it" class="PRON" usas="Z8"&gt;it&lt;/w&gt; &lt;w pos="VBZ" hw="be" class="VERB" usas="A3"&gt;is&lt;/w&gt; &lt;w pos="RG" hw="very" class= "ADV" usas="A13:3"&gt;very&lt;/w&gt; &lt;w pos="JJ" hw="hard" class=</pre>

1. Prepare your data in a folder.
2. On the 'My data' tab provide information about the corpus and navigate to the data (individual files or folders with subfolders). You can also drag and drop data into the box.

### Add corpora

Corpus hub
My data
Web

Corpus full name\*

Short display name

Language  ▼

Data folder\* i  Browse

Tagging  Grammatical i  Semantic i

Load corpus
Close

3. You can also automatically annotate (tag) corpus for pos, headword, grammatical relation and semantic (USAS) category.
4. Click on 'Load corpus'.
5. Once the corpus is loaded, click on 'Continue'

---

## 2.3 Web

#LancsBox allows you to easily scrape data from the web and create your own corpus.

---

1. On the 'Web' tab provide information about the corpus you want to create (name, language).
2. Paste a list of URLs, which you want to scrape at depth 1.
3. Decide on the additional parameter or leave defaults.
4. Click on 'Create corpus'.
5. Once the corpus is created, click on 'Continue'

---

### Add corpora

Corpus hub    My data    **Web**

Corpus full name\* \_\_\_\_\_

Short display name \_\_\_\_\_

Language    English ▼

Initial URLs\* ?

Limits    Pages ? 100 ▲▼    Depth ? 2 ▲▼

Follow external links ?

Randomize order ?

Content selector ?     Whole body     Selector ? p, h1, h2, h3, h4

Scrape links from ?     Whole body     Selector ? p, h1, h2, h3, h4


Create corpus    Close

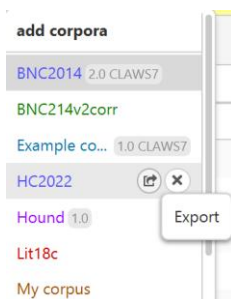
---

## 2.4 Exporting corpora

#LancsBox allows you to export corpora in XML. This functionality is available for corpora with unrestricted access.

---

Hover your mouse over the name of a corpus and click on the 'Export'  icon.





### 3 KWIC tool (Key Word In Context)

The KWIC tool generates a list of all instances of a search term in a corpus in the form of a concordance. It can be used, for example, to:

- Find the frequency of a word or phrase in a corpus.
- Find frequencies of different word classes such as nouns, verbs, adjectives.
- Find complex linguistic structures such as the passives, split infinitives etc. using 'smart searches'.
- Sort concordance lines.
- Compare multiple analyses side-by-side.

#### 3.1 KWIC: An overview

The following is a simple, yet efficient design of the KWIC tool. The single search box allows users to carry out a wide variety of powerful searches.

The screenshot shows the KWIC tool interface. At the top, there is a search box containing 'cat' and a 'Save results' button. Below this, a yellow header bar displays 'BNC2014', a search box with 'magazines', and '15M'. A 'Hits: 428 (0.29)' indicator is also present. A 'Select corpus' callout points to the 'BNC2014' label. A 'Search for a word, phrase or grammatical structure' callout points to the search box. A 'Select subcorpus' callout points to the 'magazines' dropdown. The main area is a table with four columns: 'File', 'Left', 'Node', and 'Right'. A callout points to the 'Left' column header: 'Left-click column header to sort. Drag to re-arrange.' Another callout points to a '+' sign at the bottom right: 'Click the + sign to add more panels.' Below the table, the text 'Search completed.' is visible.

File	Left	Node	Right
MagT3-1...	dual - mode LTE (up to	Cat	4 at 150 Mbps). While
MagCla2...	r, but they killed that	cat	in his thirties. I soon
MagInv2...	ircassia s (CIR) novel	cat	allergy medicine failed to reduce
MagThe2...	med bay. Adventure	Cat	tours offer a day or
MagCla1...	Geezer offers reward to catch	cat	killer Black Sabbath bassist disgusted
MagCyc1...	most combative rider, two first	cat	climbs, a special prime on
MagCla1...	Convention, Nick Drake and even	Cat	Stevens, also enjoyed a certain
MagCos1...	's Binky Felstead speaks to	Cat	Sarsfield about beauty, boys and
MagCos1...	Chelsea's Lucy chats to	Cat	Sarsfield about finding her perfect
MagCla3...	was just too hard a	cat	for me. It took all
MagCos1...	win Eurovision 2014 20. A	cat	saved a little boy from
MagRev4...	their garden bushes into a	cat,	and has since created a
MagEsq9...	a traditional curse - a mutilated	cat	on the doorstep. Anger spent

Click a row in a table to select it. Hold the Ctrl or Cmd key while clicking to select multiple rows. Selected rows can be copied with the Ctrl+C / Cmd+C keyboard shortcut or by right clicking the table and selecting the "Copy" option.

Results can be also saved easily from the main menu, where 'Save' or 'Save all' can be selected to save the active panel (highlighted) or all panels respectively.

### 3.2 Multiple panels

#LancsBox X allows analyses in multiple panels. Panels can be re-arranged by clicking and dragging on the top part of the window.

Multiple panels can be selected by holding down the Ctrl or Cmd key while clicking tools. This can be used to perform the same search in multiple panels at once.

The screenshot shows the LancsBox X interface with three panels. The top-left panel is titled 'BNC2014' and shows search results for 'whole corpus' with 100M hits. The top-right panel is also titled 'BNC2014' but for 'informal speech' with 10M hits. The bottom-right panel is titled 'BNC2014' for 'academic prose' with 20M hits. Each panel displays a table with columns: File, Left, Node, and Right. Callouts point to a 'Summary of results' box in the top-right panel, a 'Table settings' box in the top-right panel, and a 'Progress bar' in the bottom-left panel.

### 3.3 Metadata columns

Efficient work with metadata is at the heart of #LancsBox X. The concordance table displays different types of meta-data. Columns can be added according to the users' need. These columns can be sorted and filtered to display relevant information. To add or remove columns in a table, click on the table settings menu (☰) and select items from the "Columns" submenu.

#LancsBox X 1.0.0

Q [word="goes" hw="go" pos="V.\*" usas="M1"]

BNC2014 whole corpus 100M

[word="goes" hw="go" pos="V.\*" usas="M1"] Hits: 13,783 (1.38) Texts: 6,894/88,171

File	Left	N...	Right	Text: ge...	Text: subgenre	Text: date ▲
FictSci85.xml	and to hide it he			fiction	fiction: sci-fi: yo...	2010
FictSci85.xml	but takes the money and			fiction	fiction: sci-fi: yo...	2010
FictMis381.xml	make money, and it all	...	directly into the company. You	fiction	fiction: miscellan...	2010
FictMis381.xml	it says that, but whoever	...	to a hundred and fifty	fiction	fiction: miscellan...	2010
FictMis399.xml	out in a passage that	...	to the surface.' I could	fiction	fiction: miscellan...	2010
FictChi13.xml	before the cart of history	...	past.' 'Come on,' Wormsley said,	fiction	fiction: children's...	2010
FictRom23.xml	No,' Dex said. 'If he	...	now, that's it, there	fiction	fiction: romance:...	2010
FictRom23.xml	him how suddenly the world	...	and changes. Here he was	fiction	fiction: romance:...	2010
FictRom23.xml	you today because when this	...	to court, you'll be	fiction	fiction: romance:...	2010
FictMis236.xml	works. Be creative: some paperwork	...	missing, one of your admirals	fiction	fiction: miscellan...	2010
FictMis390.xml	body. I decided.' This just	...	to prove you are not	fiction	fiction: miscellan...	2010
FictMys91.xml	get on there. Then it	...	to Central, where it gets	fiction	fiction: mystery: ...	2010
FictMys91.xml	plastic is best, and that	...	in one pile; blue in	fiction	fiction: mystery: ...	2010
FictMys91.xml	near me and Gardo, it	...	down the far end, and	fiction	fiction: mystery: ...	2010

Search completed.

### 3.4 Filters

Powerful filters can be applied to i) linguistic and ii) metalinguistic data. Simply hover the mouse pointer towards the right of any column header to find the filter options button ▼.

Linguistic data can be filtered using the complete linguistic search functionality. For the left and the right context, choose the position(s) where the required linguistic feature should occur.

NOUN

Matching within:

- L1
- L2
- L3
- L4
- L5
- L6
- L7
- L8

Apply Delete

Node ▼

Contains query match ▼

[pos="N.\*"]

Apply Delete

Metalinguistic data can be filtered according to three data types: i) categories, ii) numbers and iii) dates.

### Categories

Select required categories by ticking the check box next to each category or search for categories and press the select all highlighted categories button .


### Numbers

Select a range of numbers using either the min & max values or the slider.

### Dates

Select a start and end date. Dates that do not follow a valid YYYY-MM-DD pattern are displayed as categories.

## 3.5 Summary table

Data displayed as concordance lines in KWIC can also be summarised using the 'Summary table' functionality . Summary tables can be applied to both i) linguistic and ii) metalinguistic data.

- Linguistic summaries include the following pieces of information: i) hits (absolute frequency), ii) number of texts, in which the linguistic feature occurs and iii) break-down according to any other available linguistic annotation such as pos-tags, semantic tags (usas), headwords (hw) etc. representing the linguistic feature in focus.

Summary table

Q time Hits: 152,404 (15.76) Texts: 5,490/7,531

Left context: word

L1  L2  L3  L4  L5  L6  L7

Value	Hits	Texts	class	hw	pos	usas
the	26,991	3,892	2	1	2	9
this	9,621	2,493	2	1	2	4
first	8,308	2,394	1	1	1	6
same	7,637	2,387	1	1	1	2
of	6,826	2,351	1	1	3	13
a	6,633	2,314	2	1	2	9
that	4,761	1,934	2	1	3	4
some	4,459	1,916	1	1	1	5
long	4,235	1,837	2	1	3	3
in	3,560	1,669	2	1	2	11
last	2,785	1,283	3	1	4	5
every	2,171	1,223	1	1	1	2
any	2,065	1,179	2	1	2	2
from	1,890	928	2	1	3	3

Close

For example, the table above shows that at the L1 position in the concordance table the most frequent word is *the*, followed by *this*, *first*, *same*... It occurs with the absolute frequency of 26,991

at the L1 position in 3,892 different texts. In this position, *the* is tagged as two pos-tags AT and RT42 and 9 different semantic usas tags. The details about the tags and their frequencies are revealed in tooltips with the mouse-over functionality.

- Meta-data summaries show a break-down according to a selected category. They include the following pieces of information: i) size of the component, ii) hits (absolute frequency) in the component, iii) relative frequency in the component, and iv) number of texts in which the linguistic feature occurs in the component out of all texts in the component.

Summary table

Q time Hits: 152,404 (15.76) Texts: 5,490/7,531

Text: genre

Value	Size	Hits	Relative freq	Texts
formal speech	6M	11,807	19.86	690/755
fiction	16M	30,155	19.16	457/458
informal speech	4M	7,250	18.38	1,779/3,635
elanguage	209K	376	17.97	7/7
other	15M	25,963	17.07	691/741
written-to-be-spoken	1M	2,024	16.25	34/34
magazines	7M	11,428	15.58	211/211
other informative	20M	28,469	14.32	638/640
newspapers	9M	13,181	14.20	435/486
official documents	2M	2,658	13.75	58/59
academic prose	16M	19,093	11.94	490/505

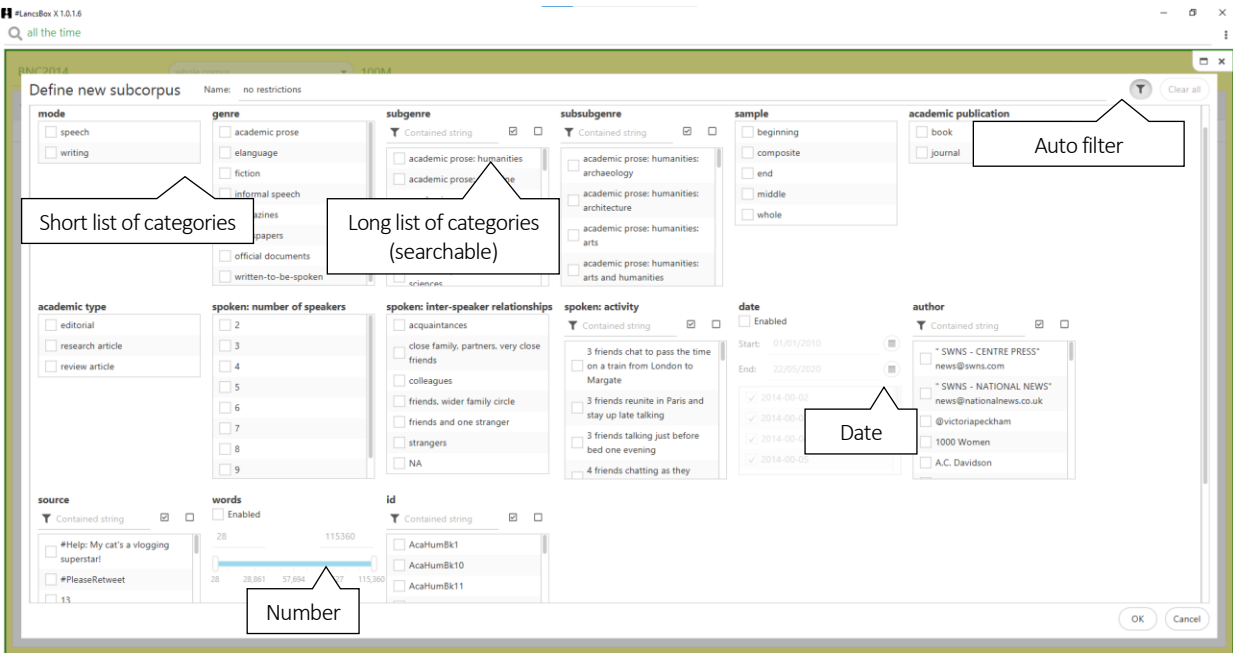
Close

Summary tables can be copied & pasted or saved; saving will also include a break-down by individual tags displayed in tooltips.

### 3.6 Working with subcorpora

#LancsBox X allows users to define subcorpora. In this way, you can restrict searches to specific parts of a corpus. To define a new subcorpus, click the subcorpus dropdown and select the “new subcorpus” option.

In the overlay that opens you can select the criteria for defining your subcorpus and choose a name. Click “OK” once all criteria have been chosen. Your new subcorpus will be selected.



You can change subcorpus using the subcorpus dropdown. The edit and delete buttons in the dropdown allow you to change or remove the subcorpora you've defined.



## 4 GraphColl

The GraphColl tool identifies collocations and displays them in a table and as a collocation graph or network.

It can be used, for example, to:

- Find the collocates of a word or phrase.
- Find colligations (co-occurrence of grammatical categories).
- Visualise collocations and colligations.
- Identify shared collocates of words or phrases.
- Summarise discourse in terms of its 'aboutness'.

### 4.1 GraphColl: An overview

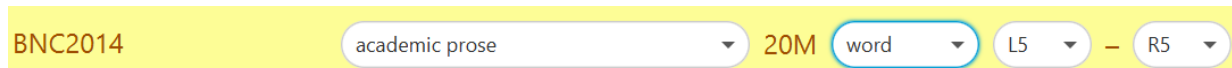
The screenshot shows the GraphColl interface for a search of the word "time". The interface includes a search bar, a "Save results" button, and a "Settings" panel. The main display is divided into two parts: a "Collocation table" on the left and a "Graph filter" network on the right. The table lists collocates with their frequencies and other statistical measures. The network graph visualizes the relationships between the search term and its collocates, with nodes representing words and edges representing their co-occurrence. A color scale at the bottom of the graph indicates the strength of the relationships, ranging from 2,000,000 to 4,000,000.

Collocate	Distrib...	Freq. (	Freq. (co...	Log	MI	Delta P1	Delta P2
at		29,034	534,122	10.4	5.1	0.2	0.05
first		11,453	112,402	10.4	6.0	0.07	0.1
same		7,019	57,294	10.0	6.3	0.04	
long		6,578	57,304	9.9	6.2	0.04	
this		17,620	496,339		4.5	0.1	0.03
for		27,6			4.3	0.2	0.03
all		10,5			4.5	0.06	0.04
last		5,474	88,027	9.5	5.3	0.03	0.06
every		4,311	44,967	9.4	5.9	0.03	0.09
the		107,956	5,258,945	9.3	3.7	0.6	0.02
it		28,683	1,283,810	9.3	3.8	0.2	0.02
time		6,492	164,522	9.3	4.6	0.04	0.04
was		18,529	843,489	9.2	3.8	0.1	0.02
a		43,406	2,210,219	9.2	3.6	0.2	0.02

## 4.2 Producing a collocation graph

GraphColl produces collocations tables and graphs on the fly. After selecting the appropriate settings you can start searching for the node and its collocates.

1. Select the appropriate settings for the collocation search:



- i) Corpus and subcorpus: Select existing or define new.
  - ii) Unit: The unit (e.g. word, headword/lemma (hw), part of speech (POS), lemma, lexeme) used for collocates.
  - iii) Span: how many words to the left (L) and to the right (R) of the node (search term) are being included in the search.
2. Type the search term into the search box (top) and press Enter.
  3. This will produce a collocation table (left) and a collocation graph (right).

## 4.3 Reading Collocation Tables

A collocation table is a traditional way of displaying collocates. In GraphColl, the table shows the following pieces of information for each collocate: i) distribution, ii) collocation frequency and iii) frequency of the collocate anywhere in the corpus, iv) all relevant statistical measures. By default, the table is sorted (largest-smallest) according to the default collocation statistic, log Dice, and an appropriate frequency filter is applied.

1. The following is a visual description of the collocation table.

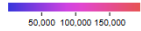
The screenshot shows a collocation table for the search term 'time'. The table has columns for Collocate, Distribution, Freq. (c), Freq. (subc...), Log D, MI, Delta P1, and Delta P2. Annotations explain various features: 'Left-click header: sort' points to the 'Log D' header; '+ Display more stats' points to a plus icon in the 'Delta P2' column; 'Right-click: assign value relevant to graph' points to the 'Freq. (subc...)' column; 'Mouse over: activate filter' points to a plus icon in the 'Delta P1' column; and 'Mouse over: KWIC preview' points to a small preview window for the collocate 'fr'.

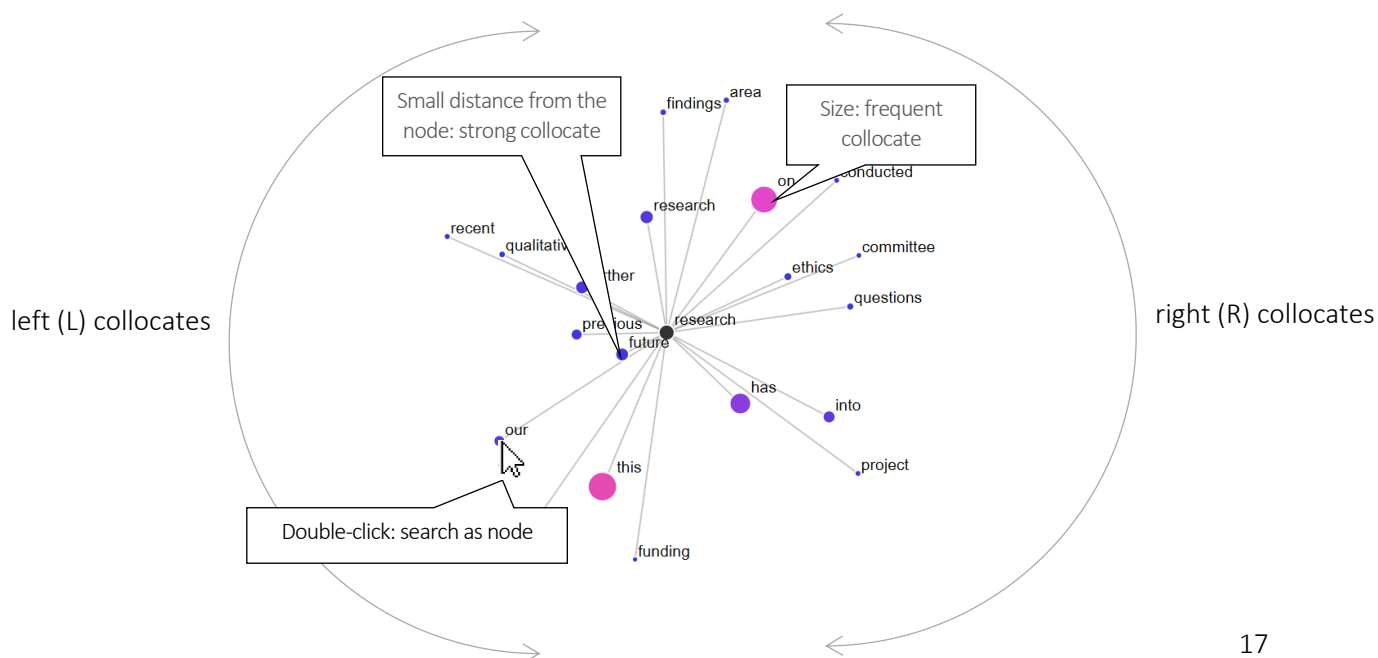
Collocate	Distribution	Freq. (c)	Freq. (subc...)	Log D	MI	Delta P1	Delta P2
at		5,984	1,051	10.9	5.9	0.2	0.08
over							
same							
period		690	7,285	9.4	6.2	0.03	0.09
first		806	17,700	9.3	5.2	0.03	0.04
spent		485	940	9.2	8.7	0.02	0.5
time		920	25,162	9.2	4.8	0.04	0.04
space		519	6,202	9.1	6.0	0.02	0.08
fr		3,629	194,288	9.1	3.9	0.1	0.02
fr		464	3,794	9.0	6.6	0.02	0.1
fr			1,250,116	9.0	3.7	0.7	0.02
fr			21,490	8.9	4.7	0.03	0.03
a		5,671	366,633	8.9	3.6	0.2	0.01
this		2,229	129,345	8.9	3.8	0.08	0.02

2. The meaning of the individual columns is:
  - i) Collocate: shows the collocate in question.
  - ii) Distribution: shows a bar chart indicating the textual position of the collocate (e.g. in the L5-R5 span).
  - iii) Freq (coll): displays the frequency of the collocation (combination of node + collocate).
  - iv) Freq (corpus): displays the frequency of the collocate anywhere in the corpus.
  - v) Stats (names): displays the values of the selected association measures; all available measures are computed at once. To display more or fewer click on the '+' button.

#### 4.4 Reading collocation graph

The graph displays multiple dimensions according to the table settings (right-click on table header to assign a graph value to a column). To find out more about a collocate, hover your mouse over it to obtain concordance lines (KWIC preview), in which the collocates co-occurs with the node.

1. Edge length: By default, the edge (line) length is assigned to a default association measure to express the strength of collocation. The closer the collocate is to the node, the stronger the association between the node and the collocate ('magnet effect').
2. Size: The size of each collocate circle is by default assigned to frequency of the collocation value: Freq (coll). The more frequent the collocation is the larger the circle.
3. Colour: The colour of each circle is by default assigned to the frequency of the collocate anywhere in the corpus: Freq (corpus). The frequency range is displayed in the legend. 
4. Position: The position of collocates around the node in the graph reflects the exact position of the collocates in text: some collocates appear (predominantly) to the left of the node, others to the right; others appear to the left and right at a similar frequency (middle position in the graph). For ease of display, if multiple collocates appear in a similar position and overlap, the tool 'spreads out' the collocates slightly.





### 4.6 Shared collocates

Shared collocates are collocates shared by at least two nodes in a graph. Shared collocates are displayed in the middle of the graph with links to the relevant nodes.

1. A full list of shared collocates can be obtained by clicking on the 'i' icon .
2. The list of shard collocates is displayed in a tabular form.

Shared collocates

Total: 344

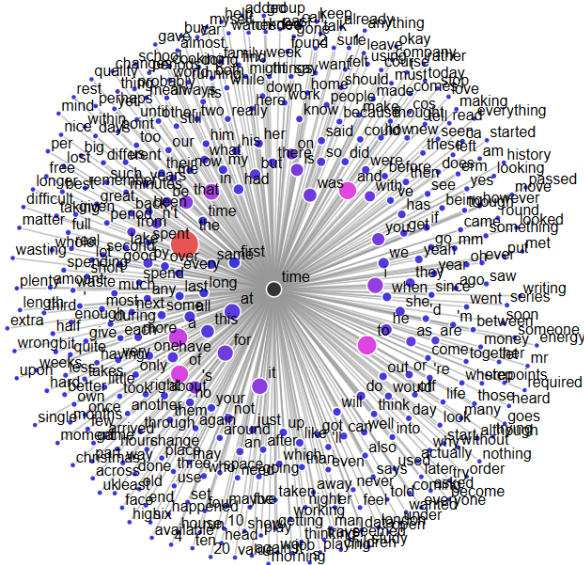
Collocate	No. of nodes ▼	Subcorpus frequency	Collocation frequencies	
			study	research
been	2	38,707	508	541
areas	2	6,175	101	120
setting	2	2,120	71	40
these	2	49,621	415	405
approved	2	540	116	70
would	2	25,125	181	195
outcomes	2	3,833	108	67
evaluation	2	1,727	162	204

Close

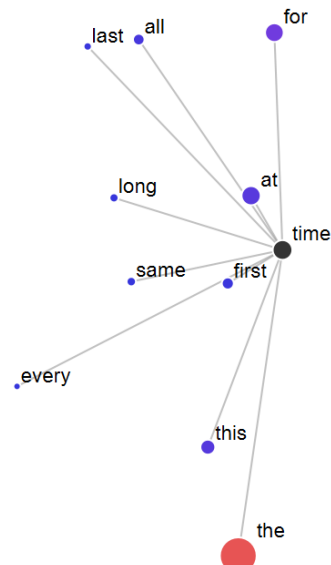
## 4.7 Problems with graphs: overpopulated graphs

If a collocation graph or network includes too many nodes and collocates, it becomes difficult to interpret. This is referred to as an overpopulated graph/network. The solution is either to change the filters in the table and make the threshold values more restrictive or to apply a filter to the graph.

The following figure shows an overpopulated graph on the left and a graph that is more easily interpretable on the right.



A graph with 392 collocates



A graph with the top 10 collocates

Choose the maximum number of collocates to show from each query. They will be selected by edge length variable.

Non-shared collocates per query

Shared collocates per query



## 4.8 Reporting collocates: CPN

It is important to realise that there is no one definite sets of collocates: different statistical procedures and threshold values highlight different sets of collocates. We therefore need to report the statistical choices involved in the identification of collocations using standard notation called Collocation Parameters Notation (CPN). When saving the results, GraphColl saves the settings in the form of CPN.

Brezina et al. (2015) propose CPN as a specific notation to be used for accurate description of collocation procedure and replication of the results. The following parameters are reported:

Statistic ID	Statistic name	Statistic cut-off value	L and R span	Minimum collocate freq. (C)	Minimum collocation freq. (NC)	Filter
4b	MI2	3	L5-R5	5	1	Function words removed
4b-MI2(3), L5-R5, C5-NC1; function words removed						

### ► Did you know?

The name GraphColl is an acronym for *graphical collocations* tool. GraphColl was the first module in #LancsBox (v.1.0) with the other tools being added at a later stage. Graphical display of collocations and collocation networks is inspired by the work of Phillips (1985), who demonstrated the concept of lexical networks (Phillip's term for 'collocation networks') with small specialised corpora. GraphColl takes this notion further, offering different statistical choices and producing collocation networks on the fly with both small and large corpora.

Phillips, M. (1985). *Aspects of text structure: An investigation of the lexical organisation of text*. Amsterdam: North-Holland.

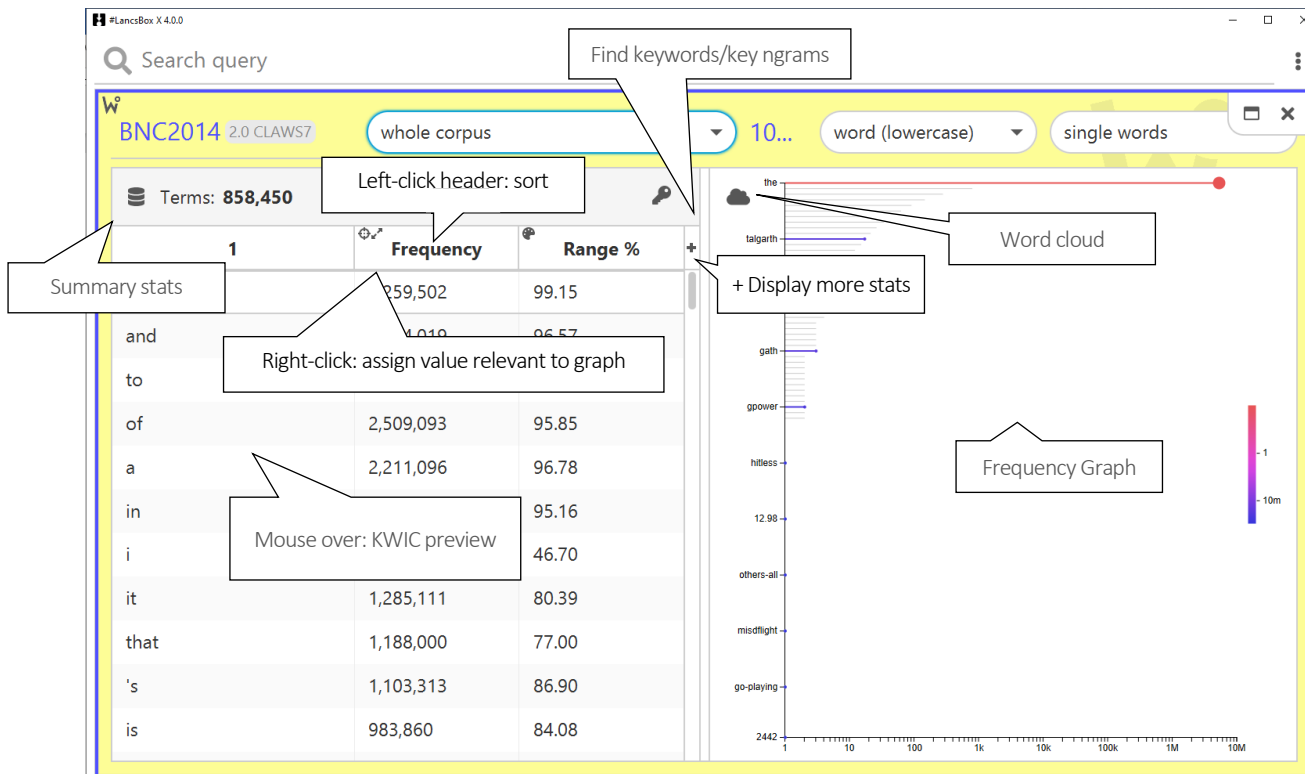
## 5 Words tool

The Words tool allows in-depth analysis of frequencies of words, n-grams, skip-grams, grammatical and semantic categories, as well as comparisons of corpora using the keywords technique.

It can be used, for example, to:

- Compute frequency and dispersion measures.
- Visualize frequency and dispersion in corpora.
- Compare corpora using the keyword technique.

### 5.1 Words: Overview



Changed corpus to BNC2014.

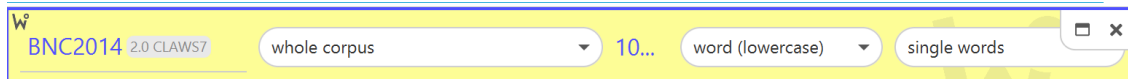
**Left:** Creating frequency lists, computing dispersion and keywords.

**Right:** Visualizing frequencies

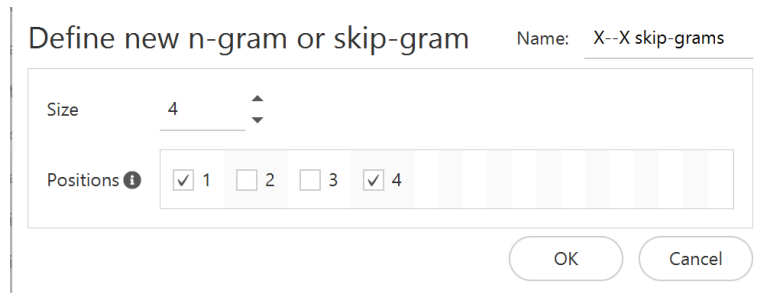
## 5.2 Producing frequency lists

When the tool is opened, Words displays a frequency list (table) based on the default corpus and default settings. These settings can be changed easily to produce different frequency lists.

1. The following are the settings for frequency lists:



- i) Corpus and subcorpus: Select existing or define new.
- ii) Unit: The unit (e.g. word, headword/lemma (hw), part of speech (POS), lemma, lexeme) used for the frequency list.
- iii) Unit size: single words, 2-grams, 3-grams, 4-grams etc., and custom n-grams and skip-grams.




2. All frequency and dispersion measures are computed at once.
3. Frequency lists can be searched using the search box (top).
4. Frequency lists can be sorted by left-clicking on any column header.
5. Frequency lists can be filtered by applying a filter to a column.



**Note:** Please note that Frequency lists in #LancsBox X are pre-computed and stored for later use. If you are creating a wordlist for the first time, this might take some time depending on the size of the corpus and complexity of its annotation (number of units used).

## 5.3 Producing keywords and key n-grams

The Words module computes a comparison of frequencies between two corpora/wordlists using a selected statistical measure.

1. Click on the key icon at the top right corner of the table .
2. Select the appropriate reference corpus.



### ► Did you know?

The statistical technique of keyword analysis was originally developed by Mike Scott (1997) and it was implemented in WordSmith Tools. It relied on corpus comparison using the chi-squared test or the log-likelihood test. As Kilgarriff pointed out, the chi-squared test and the log-likelihood test are not entirely appropriate for this type of comparison. Kilgarriff's solution implemented in Sketch Engine was to compare corpora using a 'simple maths' procedure, a simple ratio between relative frequencies of words in the two corpora we compare. In addition to 'simple maths', #LancsBox offers also other types of solutions for corpus comparison.

Scott, M. (1997). PC analysis of key words—and key key words. *System*, 25(2), 233-245.

Kilgarriff, A. (2009, July). Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference*. Liverpool, UK.

## 6 Text tool

The Text tool provides an overview of all files (texts) in the corpus, their size and lexical diversity. It also allows in-depth analysis of individual texts in the full view mode. The tool also searches texts and offer an overview table with a breakdown of frequencies and relative frequencies per file. The tool also highlights search terms in individual texts.

It can be used, for example, to:

- Explore corpora and their files (texts) before analysing them.
- Visualize corpus files and understand their distribution in terms of their sizes, lexical diversity and frequencies of linguistic features in them.
- Qualitatively analyse texts.

### 6.1 Text: Overview

The screenshot shows the Text tool interface with the following components and annotations:

- Search query:** A search bar at the top with the text "Search query".
- Number of files in the (sub)corpus:** A dropdown menu showing "88,171" files.
- Table:** A table with columns: Name, Tokens, MATTR<sub>50</sub>, and MTL. The table lists several files with their respective values.
- Graph filter:** A filter icon (funnel) above the graph.
- Graph showing individual files:** A bubble chart where the size of each bubble represents a file's characteristics.
- Annotations:** Callouts explain various actions: "Search in individual texts", "View annotation", "Right-click: assign value relevant to graph", and "Left-click: sort".

Name	Tokens	MATTR <sub>50</sub>	MTL
AcaSocBk7.xml	31,107	0.83	102.49
FictMis425.xml	29,701	0.83	123.33
ElanSocFac11.xml	29,047	0.86	128.17
FictHis11.xml	27,120	0.80	77.89
Sp2m6f1.xml	25,841	0.71	38.83
AcaPleBk22.xml	25,533	0.82	107.25
Sp2m2f25.xml	25,416	0.74	46.30

Changed subcorpus to whole corpus.

**Left:** Overview table or full text view.

**Right:** Visualizing corpus files



#LancsBox X 3.0.0

new

BNC2014 2.0 CLAWS7 whole corpus 102M

new Hits: 96,237 (940.68) Texts: 28,976/88,171

Mis397.xml 3 (571.43)

Tokens	MATTR <sub>50</sub>	MTLD	mode
5,250	0.79	66.56	writing

↑ ↓ Move to the next occurrence of the search term

stones - the new stones: they've never seemed the same to me as the old ones - and there was Carine inside Circle looking sort of tired and shrunken and old. Mitch was beside her, the last one of Tommy and Gela's grandchildren, with his scolded hair, and his blind eyes, and his hands that grabbed and groped around him all the time, like he was frightened he was sinking into the earth. Just outside Circle, Secret Ree stooped over her bits of bark.

Paaaaarp! Paaaaarp! ————— horns because the most important people hadn't yet arrived and the meeting couldn't start without them.

Relative frequencies visualized (colour)

0 500,000 1,000,000

Searched KWIC for "new".

## 7 Searching in #LancsBox

#LancsBox offers powerful searches at different levels of corpus annotation using i) simple searches, ii) wildcard searches, iii) smart searches, iv) CQL searches.

1. Simple searches are literal searches for a particular word (*new*) or phrase (*New York Times*). Simple searches are case insensitive; this means that *new*, *New*, *NEW*, *NeW* etc. will return the same set of results.
2. Wildcard searches are searches including asterisk \* as a special character.

Special character	Meaning	Example of use
*	0 or more characters	new* [ <i>new, news, newly, newspaper...</i> ]
	any word [with space]	new * [ <i>new car, New York, new ideas...</i> ]

3. Punctuation searches:

To search for punctuation use forward slashes as in the examples below.

/?/

hello /,/

4. Smart searches are searches predefined in the tool to offer users easy access to complex searches; smart searches are unique to #LancsBox. These searches are used for searching for word classes (NOUN, VERB etc.), complex grammatical patterns (PASSIVE, SPLIT\_INFINITIVE etc.) and semantic categories (PLACE\_ADVERB).

The following smart searches are available for English:

ADJECTIVE
ADVERB
BE
BODY
BOOSTER
COLLECTIVE_NOUN
COLOUR
COMPARATIVE
COMPLEX_NOUN_PHRASE
CONDITIONAL
CONNECTOR
CONTRACTION
DEGREE ADVERB
DETERMINER
DO
DOWNTONER
EMOTION

EMOTION
EXISTENTIAL_THERE
FEMALE
FEMALE
FOOD
GERUND
HAVE
HYPHENATED_WORD
INDEFINITE_PRONOUN
INFINITIVE
INFINITIVE
INTERJECTION
LINKING_ADVERB
LONG_WORD
MALE
MALE
MEDIA
MODAL

NEGATION
NOMINALIZATION
NOUN
NUMBER
PARTICLE
PASSIVE
PAST_PARTICIPLE
PAST_TENSE
PEOPLE
PEOPLE
PERFECT_INFINITIVE
PHRASAL_VERB
PLACE_ADVERB
PLANET
PREPOSITIONAL_PHRASE
PRESENT_TENSE
PRONOUN
PROPER_NOUN
REFLEXIVE_PRONOUN
SHORT_WORD
SPLIT_INFINITIVE
SUPERLATIVE
SUPERNATURAL
SUPERNATURAL
SWEARWORDS
TECHNOLOGY
TIME
TIME_ADVERB
VERB

5. CQL (Corpus Query Language searches. #LancsBox supports powerful searches using CQL.

These can be used for defining complex searches at different levels of annotation.

The levels of annotation and syntax depend on the tagging of the corpus, but for XML corpora it is common to have i) word, ii) headword/lemma (hw), iii) part-of-speech (POS), and iv) a user-defined tag. For example, a single token can be searched in CQL with

```
[word="goes" hw="go" pos="V.*" sem="M1"]
```

This will match every instance of the word *goes* with the headword *go*, the part-of-speech tag *V.\** (verb) and the *usas* tag *M1* (Moving, coming and going). If a level of annotation is not specified, no restriction is applied at that level. Everything in double quotes is interpreted as a case insensitive regular expression.

To make queries case sensitive use double equals as in the example below:

```
[word=="US"]
```

To make negative searches use a combination of an exclamation mark and the equals sign, which means 'is not equal to' as in the example below:

```
[word!="new"]
```

To search for punctuation use forward slashes and the attribute *punc* as in the example below. Note that special characters such as the question mark or the full stop need to be escaped by using the backslash symbol \

```
/punc="\?|\.|,|;"/
```

Multiple tokens can be placed in sequence. An empty pair of square brackets [] will match any token. Tokens can be repeated X times using the syntax {X}, and repeated anywhere between Y and Z times using the syntax {Y, Z}. The shorthand for {0, 1} is a question mark. Thus, for instance, the following CQL expression

```
[pos="VB.*"] []{0,3} [pos="V.N"]?
```

is interpreted as a verb to be (*VB.\**) followed by between 0 and 3 tokens without restriction (*[] {0,3}*) and optionally followed by the past participle (*V.N*).

Parts of a query can also be wrapped in parentheses (), allowing a quantifier such as {1,2} to apply to sequence of tokens—e.g. ([pos="N.\* "][word="and"]){2}. Words, phrases and smart searches can be used anywhere CQL tokens can—e.g. very{2} ADJECTIVE{1,2} [hw="year"].

CQL also supports searching XML structure. This search matches every <u></u> element, representing utterances: <u/>. The following matches every utterance where the n attribute is 1 and the nationality attribute is British or American:

```
<u n="1" nationality="British|American"/>
```

These element queries can be combined with the other types of queries using the *within* syntax:

```
[pos="D.* "] green NOUN within <text genre="newspapers"/>
```

This query matches every instance of a determiner followed by “green” followed by a noun within newspaper texts. The left and right hand sides of the *within* query can be anything; they can also be other within queries:

```
(<emoji/> within please) within (<e/> within <text genre="elanguage"/>)
```

## 8 spaCy POS tagset: English

<b>CC</b>	conjunction, coordinating	<b>PRP\$</b>	pronoun, possessive
<b>CD</b>	cardinal number	<b>RB</b>	adverb
<b>DT</b>	determiner	<b>RBR</b>	adverb, comparative
<b>EX</b>	existential there	<b>RBS</b>	adverb, superlative
<b>FW</b>	foreign word	<b>RP</b>	adverb, particle
<b>IN</b>	conjunction, subordinating or preposition	<b>SYM</b>	symbol
<b>JJ</b>	adjective	<b>TO</b>	infinitival to
<b>JJR</b>	adjective, comparative	<b>UH</b>	interjection
<b>JJS</b>	adjective, superlative	<b>VB</b>	verb, base form
<b>LS</b>	list item marker	<b>VBZ</b>	verb, 3rd person singular present
<b>MD</b>	verb, modal auxiliary	<b>VBP</b>	verb, non-3rd person singular present
<b>NNNDENCY TAG</b>	noun, singular or mass	<b>VBD</b>	verb, past tense
<b>NNS</b>	noun, plural	<b>VBN</b>	verb, past participle
<b>NNP</b>	noun, proper singular	<b>VBG</b>	verb, gerund or present participle
<b>NNPS</b>	noun, proper plural	<b>WDT</b>	<i>wh</i> -determiner
<b>PDT</b>	predeterminer	<b>WP</b>	<i>wh</i> -pronoun, personal
<b>POS</b>	possessive ending	<b>WP\$</b>	<i>wh</i> -pronoun, possessive
<b>PRP</b>	pronoun, personal	<b>WRB</b>	<i>wh</i> -adverb



## 9 spaCy dependency tags

acl	clausal modifier of noun (adjectival clause)
acomp	adjectival complement
advcl	adverbial clause modifier
advmod	adverbial modifier
agent	agent
amod	adjectival modifier
appos	appositional modifier
attr	attribute
aux	auxiliary
auxpass	auxiliary (passive)
case	case marking
cc	coordinating conjunction
ccomp	clausal complement
compound	compound
conj	conjunct
csubj	clausal subject
csubjpass	clausal subject (passive)
dative	dative
dep	unclassified dependent
det	determiner
dobj	direct object
expl	expletive
intj	interjection
mark	marker
meta	meta modifier
neg	negation modifier
nmod	modifier of nominal
npadvmod	noun phrase as adverbial modifier
nsubj	nominal subject
nsubjpass	nominal subject (passive)
nummod	numeric modifier
oprd	object predicate
parataxis	parataxis
pcomp	complement of preposition
pobj	object of preposition
poss	possession modifier
preconj	pre-correlative conjunction
predet	None
prep	prepositional modifier
prt	particle
punct	punctuation
quantmod	modifier of quantifier
relcl	relative clause modifier
xcomp	open clausal complement

## 10 CLAWS tagset (C7)

Source: <http://ucrel.lancs.ac.uk/claws7tags.html>

<b>APPG</b>	possessive pronoun, pre-nominal (e.g. my, your, our)
<b>AT</b>	article (e.g. the, no)
<b>AT1</b>	singular article (e.g. a, an, every)
<b>BCL</b>	before-clause marker (e.g. in order (that), in order (to))
<b>CC</b>	coordinating conjunction (e.g. and, or)
<b>CCB</b>	adversative coordinating conjunction ( but)
<b>CS</b>	subordinating conjunction (e.g. if, because, unless, so, for)
<b>CSA</b>	as (as conjunction)
<b>CSN</b>	than (as conjunction)
<b>CST</b>	that (as conjunction)
<b>CSW</b>	whether (as conjunction)
<b>DA</b>	after-determiner or post-determiner capable of pronominal function (e.g. such, former, same)
<b>DA1</b>	singular after-determiner (e.g. little, much)
<b>DA2</b>	plural after-determiner (e.g. few, several, many)
<b>DAR</b>	comparative after-determiner (e.g. more, less, fewer)
<b>DAT</b>	superlative after-determiner (e.g. most, least, fewest)
<b>DB</b>	before determiner or pre-determiner capable of pronominal function (all, half)
<b>DB2</b>	plural before-determiner ( both)
<b>DD</b>	determiner (capable of pronominal function) (e.g. any, some)
<b>DD1</b>	singular determiner (e.g. this, that, another)
<b>DD2</b>	plural determiner ( these, those)
<b>DDQ</b>	wh-determiner (which, what)
<b>DDQGE</b>	wh-determiner, genitive (whose)
<b>DDQV</b>	wh-ever determiner, (whichever, whatever)
<b>EX</b>	existential there
<b>FO</b>	formula
<b>FU</b>	unclassified word
<b>FW</b>	foreign word
<b>GE</b>	germanic genitive marker - (' or 's)
<b>IF</b>	for (as preposition)
<b>II</b>	general preposition
<b>IO</b>	of (as preposition)
<b>IW</b>	with, without (as prepositions)

**JJ** general adjective  
**JJR** general comparative adjective (e.g. older, better, stronger)  
**JJT** general superlative adjective (e.g. oldest, best, strongest)  
**JK** catenative adjective (able in be able to, willing in be willing to)  
**MC** cardinal number, neutral for number (two, three..)  
**MC1** singular cardinal number (one)  
**MC2** plural cardinal number (e.g. sixes, sevens)  
**MCGE** genitive cardinal number, neutral for number (two's, 100's)  
**MCMC** hyphenated number (40-50, 1770-1827)  
**MD** ordinal number (e.g. first, second, next, last)  
**MF** fraction, neutral for number (e.g. quarters, two-thirds)  
**ND1** singular noun of direction (e.g. north, southeast)  
**NN** common noun, neutral for number (e.g. sheep, cod, headquarters)  
**NN1** singular common noun (e.g. book, girl)  
**NN2** plural common noun (e.g. books, girls)  
**NNA** following noun of title (e.g. M.A.)  
**NNB** preceding noun of title (e.g. Mr., Prof.)  
**NNL1** singular locative noun (e.g. Island, Street)  
**NNL2** plural locative noun (e.g. Islands, Streets)  
**NNO** numeral noun, neutral for number (e.g. dozen, hundred)  
**NNO2** numeral noun, plural (e.g. hundreds, thousands)  
**NNT1** temporal noun, singular (e.g. day, week, year)  
**NNT2** temporal noun, plural (e.g. days, weeks, years)  
**NUU** unit of measurement, neutral for number (e.g. in, cc)  
**NUU1** singular unit of measurement (e.g. inch, centimetre)  
**NUU2** plural unit of measurement (e.g. ins., feet)  
**NP** proper noun, neutral for number (e.g. IBM, Andes)  
**NP1** singular proper noun (e.g. London, Jane, Frederick)  
**NP2** plural proper noun (e.g. Browns, Reagans, Koreas)  
**NPD1** singular weekday noun (e.g. Sunday)  
**NPD2** plural weekday noun (e.g. Sundays)  
**NPM1** singular month noun (e.g. October)  
**NPM2** plural month noun (e.g. Octobers)  
**PN** indefinite pronoun, neutral for number (none)  
**PN1** indefinite pronoun, singular (e.g. anyone, everything, nobody, one)  
**PNQO** objective wh-pronoun (whom)  
**PNQS** subjective wh-pronoun (who)  
**PNQV** wh-ever pronoun (whoever)

**PNX1** reflexive indefinite pronoun (oneself)  
**PPGE** nominal possessive personal pronoun (e.g. mine, yours)  
**PPH1** 3rd person sing. neuter personal pronoun (it)  
**PPHO1** 3rd person sing. objective personal pronoun (him, her)  
**PPHO2** 3rd person plural objective personal pronoun (them)  
**PPHS1** 3rd person sing. subjective personal pronoun (he, she)  
**PPHS2** 3rd person plural subjective personal pronoun (they)  
**PPIO1** 1st person sing. objective personal pronoun (me)  
**PPIO2** 1st person plural objective personal pronoun (us)  
**PPIS1** 1st person sing. subjective personal pronoun (I)  
**PPIS2** 1st person plural subjective personal pronoun (we)  
**PPX1** singular reflexive personal pronoun (e.g. yourself, itself)  
**PPX2** plural reflexive personal pronoun (e.g. yourselves, themselves)  
**PPY** 2nd person personal pronoun (you)  
**RA** adverb, after nominal head (e.g. else, galore)  
**REX** adverb introducing appositional constructions (namely, e.g.)  
**RG** degree adverb (very, so, too)  
**RGQ** wh- degree adverb (how)  
**RGQV** wh-ever degree adverb (however)  
**RGR** comparative degree adverb (more, less)  
**RGT** superlative degree adverb (most, least)  
**RL** locative adverb (e.g. alongside, forward)  
**RP** prep. adverb, particle (e.g. about, in)  
**RPK** prep. adv., catenative (about in be about to)  
**RR** general adverb  
**RRQ** wh- general adverb (where, when, why, how)  
**RRQV** wh-ever general adverb (wherever, whenever)  
**RRR** comparative general adverb (e.g. better, longer)  
**RRT** superlative general adverb (e.g. best, longest)  
**RT** quasi-nominal adverb of time (e.g. now, tomorrow)  
**TO** infinitive marker (to)  
**UH** interjection (e.g. oh, yes, um)  
**VB0** be, base form (finite i.e. imperative, subjunctive)  
**VBDR** were  
**VBDZ** was  
**VBG** being  
**VBI** be, infinitive (To be or not... It will be ..)  
**VBM** am

<b>VCN</b>	been
<b>VBR</b>	are
<b>VBZ</b>	is
<b>VDO</b>	do, base form (finite)
<b>VDD</b>	did
<b>VDG</b>	doing
<b>VDI</b>	do, infinitive (I may do... To do...)
<b>VDN</b>	done
<b>VDZ</b>	does
<b>VH0</b>	have, base form (finite)
<b>VHD</b>	had (past tense)
<b>VHG</b>	having
<b>VHI</b>	have, infinitive
<b>VHN</b>	had (past participle)
<b>VHZ</b>	has
<b>VM</b>	modal auxiliary (can, will, would, etc.)
<b>VMK</b>	modal catenative (ought, used)
<b>VV0</b>	base form of lexical verb (e.g. give, work)
<b>VVD</b>	past tense of lexical verb (e.g. gave, worked)
<b>VVG</b>	-ing participle of lexical verb (e.g. giving, working)
<b>VVGK</b>	-ing participle catenative (going in be going to)
<b>VVI</b>	infinitive (e.g. to give... It will work...)
<b>VVN</b>	past participle of lexical verb (e.g. given, worked)
<b>VVNK</b>	past participle catenative (e.g. bound in be bound to)
<b>VVZ</b>	-s form of lexical verb (e.g. gives, works)
<b>XX</b>	not, n't
<b>ZZ1</b>	singular letter of the alphabet (e.g. A,b)
<b>ZZ2</b>	plural letter of the alphabet (e.g. A's, b's)

## 11 USAS semantic tagset

Source: <http://ucrel.lancs.ac.uk/usas>

A1	GENERAL AND ABSTRACT TERMS	A7	Definite (+ modals)	E3	Calm/Violent/Angry
A1.1.1	General actions, making etc.	A8	Seem	E4	Happy/sad
A1.1.2	Damaging and destroying	A9	Getting and giving; possession	E4.1	Happy/sad: Happy
A1.2	Suitability	A10	Open/closed; Hiding/Hidden; Finding; Showing	E4.2	Happy/sad: Contentment
A1.3	Caution	A11	Importance	E5	Fear/bravery/shock
A1.4	Chance, luck	A11.1	Importance: Important	E6	Worry, concern, confident
A1.5	Use	A11.2	Importance: Noticeability	F1	Food
A1.5.1	Using	A12	Easy/difficult	F2	Drinks
A1.5.2	Usefulness	A13	Degree	F3	Cigarettes and drugs
A1.6	Physical/mental	A13.1	Degree: Non-specific	F4	Farming & Horticulture
A1.7	Constraint	A13.2	Degree: Maximizers	G1	Government, Politics and elections
A1.8	Inclusion/Exclusion	A13.3	Degree: Boosters	G1.1	Government etc.
A1.9	Avoiding	A13.4	Degree: Approximators	G1.2	Politics
A2	Affect	A13.5	Degree: Compromisers	G2	Crime, law and order
A2.1	Affect:- Modify, change	A13.6	Degree: Diminishers	G2.1	Crime, law and order: Law and order
A2.2	Affect:- Cause/Connected	A13.7	Degree: Minimizers	G2.2	General ethics
A3	Being	A14		G3	Warfare, defence and the army; weapons
A4	Classification		Exclusivizers/particulari zers	H1	Architecture and kinds of houses and buildings
A4.1	Generally kinds, groups, examples	A15	Safety/Danger	H2	Parts of buildings
A4.2	Particular/general; detail	B1	Anatomy and physiology	H3	Areas around or near houses
A5	Evaluation	B2	Health and disease	H4	Residence
A5.1	Evaluation:- Good/bad	B3	medicines and medical treatment	H5	Furniture and household fittings
A5.2	Evaluation:- True/false	B4	Cleaning and personal care	I1	Money generally
A5.3	Evaluation:- Accuracy	B5	Clothes and personal belongings	I1.1	Money: Affluence
A5.4	Evaluation:- Authenticity	C1	Arts and crafts	I1.2	Money: Debts
A6	Comparing	E1	EMOTIONAL ACTIONS, STATES AND PROCESSES	I1.3	Money: Price
A6.1	Comparing:- Similar/different		General	I2	Business
A6.2	Comparing:- Usual/unusual	E2	Liking	I2.1	Business: Generally
A6.3	Comparing:- Variety			I2.2	Business: Selling
				I3	Work and employment

I3.1	Work and employment: Generally	N3.8	Measurement: Speed	Q4.2	The Media:- Newspapers etc.
I3.2	Work and employemny: Professionalism	N4	Linear order	Q4.3	The Media:- TV, Radio and Cinema
I4	Industry	N5	Quantities	S1	SOCIAL ACTIONS, STATES AND PROCESSES
K1	Entertainment generally	N5.1	Entirety; maximum	S1.1	SOCIAL ACTIONS, STATES AND PROCESSES
K2	Music and related activities	N5.2	Exceeding; waste	S1.1.1	SOCIAL ACTIONS, STATES AND PROCESSES
K3	Recorded sound etc.	N6	Frequency etc.	S1.1.1.1	SOCIAL ACTIONS, STATES AND PROCESSES
K4	Drama, the theatre and showbusiness	O1	Substances and materials generally	S1.1.2	Reciprocity
K5	Sports and games generally	O1.1	Substances and materials generally: Solid	S1.1.3	Participation
K5.1	Sports	O1.2	Substances and materials generally: Liquid	S1.1.4	Deserve etc.
K5.2	Games	O1.3	Substances and materials generally: Gas	S1.2	Personality traits
K6	Childrens games and toys	O2	Objects generally	S1.2.1	Approachability and Friendliness
L1	Life and living things	O3	Electricity and electrical equipment	S1.2.2	Avarice
L2	Living creatures generally	O4	Physical attributes	S1.2.3	Egoism
L3	Plants	O4.1	General appearance and physical properties	S1.2.4	Politeness
M1	Moving, coming and going	O4.2	Judgement of appearance (pretty etc.)	S1.2.5	Toughness; strong/weak
M2	Putting, taking, pulling, pushing, transporting &c.	O4.3	Colour and colour patterns	S1.2.6	Sensible
M3	Vehicles and transport on land	O4.4	Shape	S2	People
M4	Shipping, swimming etc.	O4.5	Texture	S2.1	People:- Female
M5	Aircraft and flying	O4.6	Temperature	S2.2	People:- Male
M6	Location and direction	P1	Education in general	S3	Relationship
M7	Places	Q1	LINGUISTIC ACTIONS, STATES AND PROCESSES;	S3.1	Relationship: General
M8	Remaining/stationary	Q1.1	LINGUISTIC ACTIONS, STATES AND PROCESSES;	S3.2	Relationship: Intimate/sexual
N1	Numbers	Q1.2	Paper documents and writing	S4	Kin
N2	Mathematics	Q1.3	Telecommunications	S5	Groups and affiliation
N3	Measurement	Q2	Speech acts	S6	Obligation and necessity
N3.1	Measurement: General	Q2.1	Speech etc:- Communicative	S7	Power relationship
N3.2	Measurement: Size	Q2.2	Speech acts	S7.1	Power, organizing
N3.3	Measurement: Distance	Q3	Language, speech and grammar	S7.2	Respect
N3.4	Measurement: Volume	Q4	The Media	S7.3	Competition
N3.5	Measurement: Weight	Q4.1	The Media:- Books	S7.4	Permission
N3.6	Measurement: Area			S8	Helping/hindering
N3.7	Measurement: Length & height			S9	Religion and the supernatural
				T1	Time
				T1.1	Time: General
				T1.1.1	Time: General: Past
				T1.1.2	Time: General: Present; simultaneous

T1.1.3	Time: General: Future	X2.5	Understand	X9.1	Ability:- Ability, intelligence
T1.2	Time: Momentary	X2.6	Expect	X9.2	Ability:- Success and failure
T1.3	Time: Period	X3	Sensory	Y1	Science and technology in general
T2	Time: Beginning and ending	X3.1	Sensory:- Taste	Y2	Information technology and computing
T3	Time: Old, new and young; age	X3.2	Sensory:- Sound	Z0	Unmatched proper noun
T4	Time: Early/late	X3.3	Sensory:- Touch	Z1	Personal names
W1	The universe	X3.4	Sensory:- Sight	Z2	Geographical names
W2	Light	X3.5	Sensory:- Smell	Z3	Other proper names
W3	Geographical terms	X4	Mental object	Z4	Discourse Bin
W4	Weather	X4.1	Mental object:- Conceptual object	Z5	Grammatical bin
W5	Green issues	X4.2	Mental object:- Means, method	Z6	Negative
X1	PSYCHOLOGICAL ACTIONS, STATES AND PROCESSES	X5	Attention	Z7	If
X2	Mental actions and processes	X5.1	Attention	Z8	Pronouns etc.
X2.1	Thought, belief	X5.2	Interest/boredom/excited/energetic	Z9	Trash can
X2.2	Knowledge	X6	Deciding	Z99	Unmatched
X2.3	Learn	X7	Wanting; planning; choosing		
X2.4	Investigate, examine, test, search	X8	Trying		
		X9	Ability		



## 12 Definitions of smart searches

ADJECTIVE	[pos="J.*"]
ADVERB	[pos="R.*"]
BE	[pos="VB.*"]
BOOSTER	[hw="absolutely altogether completely enormously entirely extremely fully greatly highly intensely perfectly strongly thoroughly totally utterly very"]
COLLECTIVE_NOUN	[hw="a" pos="D.*"] [hw="aerie album ambush anthology archipelago argument argumentation armada army array arsenal ascension assembly aurora badelynge bag bale band bank banner barrel barren bask basket batch battery bazaar bed bellowing belt bench bevy bew bill bind bits blessing bloat block blush board bob body boil boll bond book bouquet bowl brace branch brew brigade brood bubble budget building bunch bundle bury business cache canteen caravan cartload cast caste catalogue catch cavalcade celebration cete chain charm chatter chattering chest chine choir chorus circle circus clamour clan clash clashing class clattering clew clique cloud clowder cluck clump cluster clutch clutter coalition coil collection colony column comb commonwealth communion community company compendium confab conflagration confraternity confusion congregation congress conspiracy constellation converting convocation convoy copse cornucopia corps cortege cost cote coterie coven cover covert covey cowardice cran crash crate creche crew crop crowd cry culture death deceit deck den descent desert destruction dicker disguising dissimulation diving division dooding dole dopping dout down doyft draft draught dray drift dropping drove drum dule durante dynasty earth eleven embarrassment equivocation erst escargatoire exaltation faculty faggot fall family farrow fellowship fesnyng fesnyng festival fesynes fidget field fine fitting fixie flange flap fleet flick flight fling flink float flock flotilla flourish flush fluther flutter fold forest fraunch fun gaggle galaxy gam gang garland garrison gathering gatling gaze generation giggle glaring gleam glide glint glitter glory glossary grist group grove gulp hail hand haras harem harvest haul head heap heard hedge herd hill hive holliness horde host house hover huddle hunt hurtle husk illusion implausibility index infestation intrusion invention kaleidoscope kettle kindle kine kingdom knab knob knot labour lamentation layer lead leap leash lepe library line list litter lodge loft lounge loveliness machination malapertness marvel mask mass match melody memory menagerie mess mews miller mischief mob mouthful movement multiply murder murmuration muscle muster mustering mutation mute necklace nest neverthriving nide nosegay nuisance number nursery nye obesiance observance obstinacy orchard orchestra ostentation outfit pace pack packet padding pair panel panes pantheon parade parcel parel park parliament party passel patrol peal peep pencil piddle pile pint pit piteousness pitying plague platoon plump pocket pod ponder pontification pool posse pounce poverty prattle pretying prickle pride prudence puddling pump punnet purse quabble quarrel quire quiver rabble radiance raffle raft rafter rag rainbow rake rangale range rayful ream reel regiment rhumba richesse ring roll romp rookery roost rope rouleau round rout route row royalty rumble rump rumpus run rush salvo sarcasm sault scatter school scold scorn scourge screech scurry sea sect sedge sequitur series serving set setting sheaf shelf shimmer shitload shoal shower shrewdness shuffle siegel singular sizzle skein skirl skulk slate sleuth slew slither sloth smack snarl snatch sneak sord sunder soviet sowse span spawn spinney spring sprinkle squad squadron stable stack staff stage stalk stand staple stare state stench stick stock storytelling streak stream string stud suit suite superfluity sute swarm swirl tassel team tenement thought threatening thunder tiding tittering toil tok torment totter tower trace train trembling tribe trimming trip troop troubling troupe truss tuft tumult turn ubiquity unkindness venue vineyard volery wad waddle wake walk warren watch wealth wedge weyr wheel whiteness whoop wing wisdom wisp wolfpack wrack wreath yap yoke zap zea zoo"] [hw="of"] [pos="NN.*"] {1,2}
COMPARATIVE	[pos="JJR RGR RRR"]
COMPLEX_NOUN_PHRASE	[pos="J.*"] {1,5} [pos="NN.*"]
CONDITIONAL	[hw="if unless"]
CONNECTOR	[pos="I.* CS CC"]
CONTRACTION	[word="s re ve d m em ll) n't" pos="^G.*"]
DEGREE_ADVERB	[hw="very really too quite exactly right pretty real more relatively" pos="R.*"]
DETERMINER	[pos="D.*"]
DO	[hw="do" pos="VV.*"]

DOWNTONER	[hw="almost barely hardly merely mildly nearly only partially partly practically scarcely slightly somewhat"]
EXISTENTIAL_THERE	[pos="EX"]
GERUND	[hw="(?!(*thing evening morning viking)).{2,}ing" pos="NN[12]"]
HAVE	[pos="VH.*"]
INFINITIVE	[pos="TO"][pos="V.*"]
HYPHENATED_WORD	[word="*.*"]
INDEFINITE_PRONOUN	[hw="anybody anyone anything everybody everyone everything nobody none nothing nowhere somebody someone something"]
INFINITIVE	[pos="TO"][pos="V.*"]
INTERJECTION	[pos="UH"]
LINKING_ADVERB	[hw="then so anyway though however e\?.?g\?.? i\?.?e\?.? therefore thus nevertheless nonetheless" pos="R.*"]
LONG_WORD	[word=".{15,}"]
MODAL	[pos="MD"]
NEGATION	[word="not .*n't no neither nowhere never nor none nobody nothing"]
NOMINALIZATION	[word=".{3,}{tion tions ment ments ness nesses ity ities}"]
NOUN	[pos="N.*"]
NUMBER	[pos="M.*"]
PARTICLE	[pos="RP"]
PASSIVE	[pos="VB[^0].*"] [pos="R.*"] {0,3} [pos="V.N"]
PAST_TENSE	[pos="V.D.?" ]
PAST_PARTICIPLE	[pos="V.N"]
PERFECT_INFINITIVE	[pos="TO"] [pos="VH.*"] [pos="V.N"]
PHRASAL_VERB	[pos="VV." ] [pos="PP.*"] {0,1} [pos="RP"]
PLACE_ADVERB	[hw="aboard above abroad across ahead alongside around ashore astern away behind below beneath beside downhill downstairs downstream east far hereabouts indoors inland inshore inside locally near nearby north nowhere outdoors outside overboard overland overseas south underfoot underneath uphill upstairs upstream west"]
PREPOSITIONAL_PHRASE	[pos="I.* CS"] [pos="J.* PP.* CC D.* RR M.* GE N.*"] {0,5} [pos="N.*"]
PRESENT_PARTICIPLE	[pos="V.GK?"]

PRESENT_TENSE	[pos="V.Z"]
PRONOUN	[pos="P.*"]
PROPER_NOUN	[pos="NP.*"]
REFLEXIVE_PRONOUN	[hw=".*sel(f ves)" pos="P.X."]
SHORT_WORD	[word=".{1,3}"]
SPLIT_INFINITIVE	[pos="TO"][pos="R.*"][pos="V.*"]
SUPERLATIVE	[pos="DAT JJT RGT RRT"]
SWEARWORDS	[hw="arse arsehole bastard bellend bint bitch bloodclaat bloody bollocks bugger bullshit clunge cock crap cunt damn dick dickhead fanny feck fuck.* gash git god goddam jesus minge minger motherfucker munter piss prick punani pussy shit sod tit twat"]
TIME_ADVERB	[hw="afterwards? again earlier early eventually formerly immediately initially instantly late lately later momentarily now nowadays once originally presently previously recently shortly simultaneously soon subsequently today tomorrow tonight yesterday"]
VERB	[pos="V.*"]
PEOPLE	[sem="S2 S2:1 S2:2 S3 S3:1 S3:2 S4"]
MALE	[sem="S2:2"]
FEMALE	[sem="S2:1"]
SUPERNATURAL	[sem="S9"]
EMOTION	[sem="E E1 E2 E3 E4 E4:1 E4:2 E5 E6"]
TIME	[sem="T1 T1:1 T1:1:1 T1:1:2 T1:2 T1:3 T2 T3 T4"]
PLANET	[sem="W1 W2 W3 W4 W5 L1 L2 L3"]
COLOR	[sem="O4:3"]
COLOUR	[sem="O4:3"]
BODY	[sem="B1 B2 B3"]
FOOD	[sem="F1 F2"]
TECHNOLOGY	[sem="Y1 Y2"]
MEDIA	[sem="Q4 Q4:1 Q4:2 Q4:3 K1 K2 K3 K4"]

## 13 Glossary

**Absolute (or raw) frequency** – The number of times a linguistic feature occurs in a corpus or its part(s); the number of hits of a search query in a corpus.

**Colligation** – Systematic co-occurrence of grammatical categories (e.g. POS tags) in text identified statistically.

**Collocate** – A word that systematically occurs with the node (word or phrase of interest, search term).

**Collocation** – Systematic co-occurrence of words in text identified statistically.

**Concordance line** – A single line in the KWIC table, usually containing the node (search match) and several words before and after it (the right and left context).

**Concordance** is a typical form of display for examples of language use found in a corpus with the node (search match) in the middle and several words of context displayed on the left and right. Concordance is sometimes also called a 'KWIC (display)'.

**Corpus** (pl. corpora) – A collection of language data that can be searched by a computer.

**Frequency** – The number of times a search query matches text in the corpus. A distinction is made between absolute (simple number of hits) and relative frequency (number of hits per X number of words).

**KWIC** – an abbreviation for 'keyword in context'. This is a typical form of display for examples found in a corpus with the node (word or phrase of interest) in the middle and several words of context displayed on the left and right. KWIC is sometimes also called a 'concordance'.

**Left context** – The words preceding a particular search match (node). Individual positions in the left-context are referred to as L1 (position immediately preceding), L2, L3 etc.

**Lemma / Headword** – All inflected forms belonging to one stem. For example, a lemma 'go' includes the following word forms (types): 'go', 'goes', 'went', 'going' and 'gone'.

**Node** – The word, phrase or grammatical structure of interest; the text matching a search query.

**Part-of-speech (POS)** – A grammatical category, a word class. Part-of-speech is usually assigned automatically using a process called part-of-speech tagging (see below).

**Part-of-speech tagging (POS tagging)** – A process of adding information about the grammatical category of each word in a text or corpus. For example, the following sentence was POS-tagged: Automatically\_RB annotates\_VBZ data\_NNS for\_IN part-of-speech\_NN.

**Regular expressions (regex)** – A special meta-language that allows advanced users to search for many strings simultaneously.

**Relative (or normalized) frequency (RF)** is calculated as the absolute frequency of a search query divided by the total number of words searched (the number of words in the corpus or subcorpus). This number is usually multiplied by an appropriate basis for normalization (e.g. 10,000).

**Right context** – The words following a particular search match (node). Individual positions in the right-context are referred to as R1 (position immediately following), R2, R3 etc.

**Subcorpus** (pl. subcorpora) – A user-defined part of a corpus which searches can be restricted to. It can include whole texts or parts of multiple texts. In #LancsBox X, subcorpora are defined using XML structure.

**Tagging** – The process of adding linguistic information to the words in a text or corpus, automatically or semi-automatically. See Part-of-speech tagging.

**Text** – A basic unit of a corpus; a corpus is a collection multiple texts.

**Token** – a single occurrence of a word form in a text or corpus.

**XML** – An abbreviation for Extensible Markup Language. A machine-readable way of writing information in text files that gives structure and annotation to the information. In corpora, XML can annotate words with part-of-speech information and give structure to texts, for example with sections and paragraphs.

Developed @ Lancaster University

