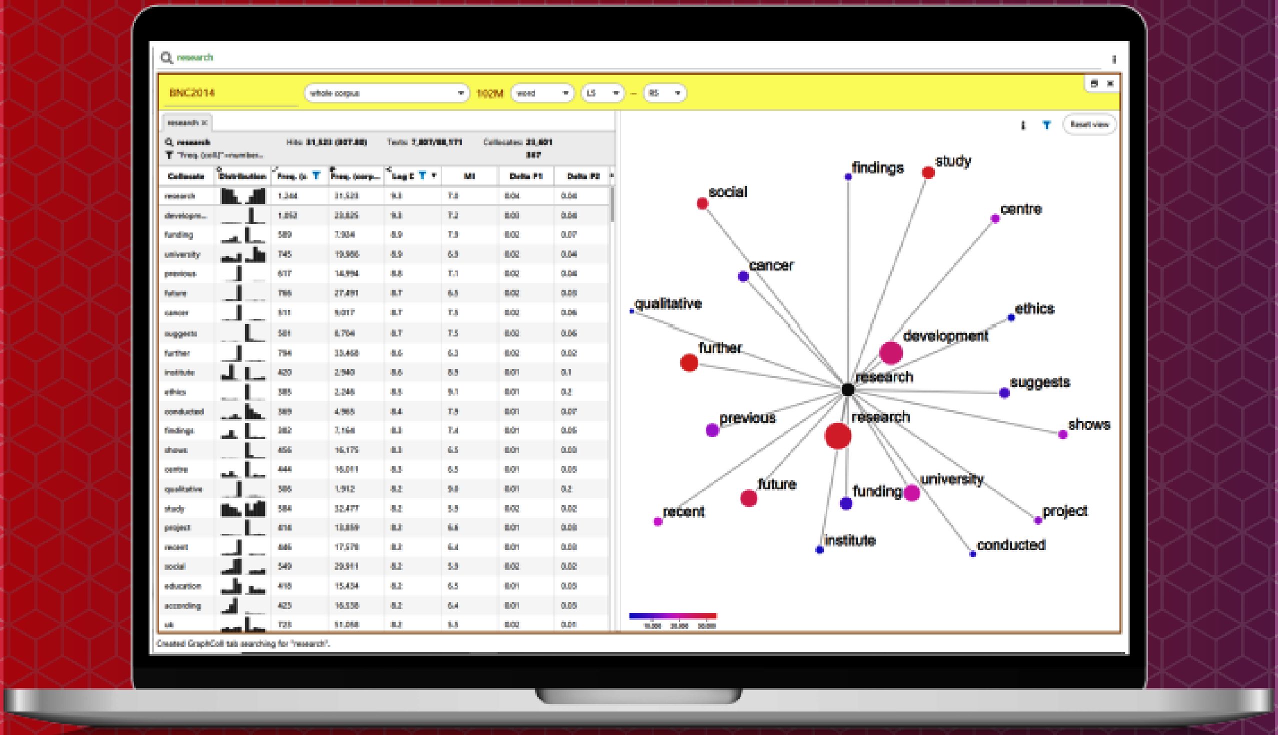
#LancsBox

Innovation in Corpus Linguistics

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

Download it for free.



research

BNC2014 | whole corpus | 102M | word | L5 – R5

research ×

research Hits: 31,523 (307.88) Texts: 7,807/88,171 Collocates: 23,601 387

"Freq. (coll.)" <number...

Collocate	Distribution	Freq. (c...	Freq. (corp...	Log £	MI	Delta P1	Delta P2
research		1,344	31,523	9.3	7.0	0.04	0.04
developm...		1,052	23,825	9.3	7.2	0.03	0.04
funding		589	7,924	8.9	7.9	0.02	0.07
university		745	19,986	8.9	6.9	0.02	0.04
previous		617	14,994	8.8	7.1	0.02	0.04
future		766	27,491	8.7	6.5	0.02	0.03
cancer		511	9,017	8.7	7.5	0.02	0.06
suggests		501	8,704	8.7	7.5	0.02	0.06
further		784	33,468	8.6	6.3	0.02	0.02
institute		420	2,940	8.6	8.9	0.01	0.1
ethics		385	2,246	8.5	9.1	0.01	0.2
conducted		389	4,985	8.4	7.9	0.01	0.07
findings		382	7,164	8.3	7.4	0.01	0.05
shows		456	16,175	8.3	6.5	0.01	0.03
centre		444	16,011	8.3	6.5	0.01	0.03
qualitative		386	1,912	8.2	9.8	0.01	0.2
study		584	32,477	8.2	5.9	0.02	0.02
project		414	13,859	8.2	6.6	0.01	0.03
recent		446	17,578	8.2	6.4	0.01	0.03
social		549	29,911	8.2	5.9	0.02	0.02
education		418	15,434	8.2	6.5	0.01	0.03
according		425	16,538	8.2	6.4	0.01	0.03
uk		723	51,058	8.2	5.5	0.02	0.01

Created GraphColl tab searching for 'research'.

Brezina, V., Platt, W. (2024). #LancsBox X 5.0.0 [software package], 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

Developed @ Lancaster University

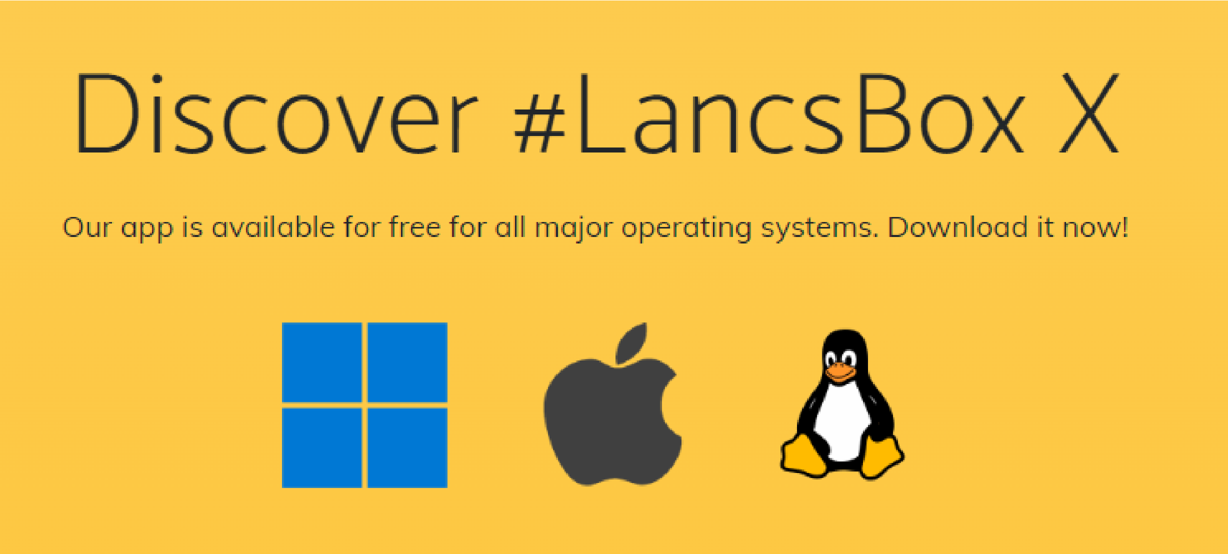
#LancsBox X: License

#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.

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



Discover #LancsBox X

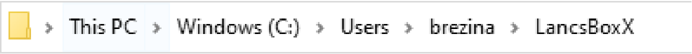
Our app is available for free for all major operating systems. Download it now!

Or simply click on

❷ **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 

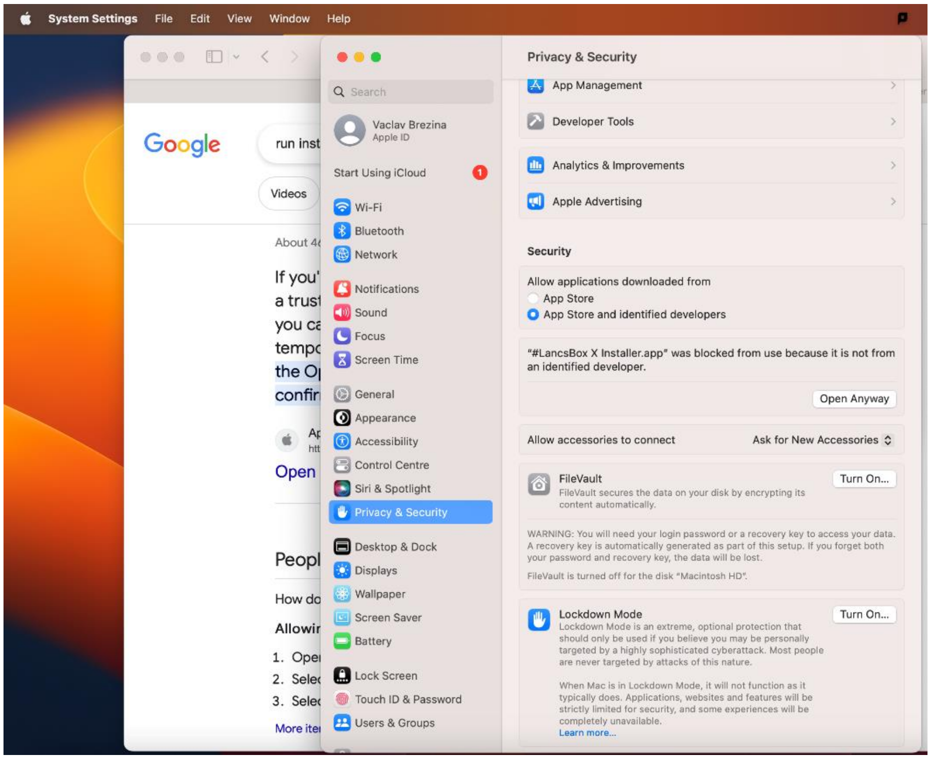
› This PC › Windows (C:) › Users › brezina › LancsBoxX

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’.



Privacy & Security

App Management

Developer Tools

Analytics & Improvements

Apple Advertising

Security

Allow applications downloaded from

○ App Store

● App Store and identified developers

“#LancsBox X Installer.app” was blocked from use because it is not from an identified developer.

Open Anyway

Allow accessories to connect: Ask for New Accessories

FileVault — Turn On...

FileVault secures the data on your disk by encrypting its content automatically.

WARNING: You will need your login password or a recovery key to access your data. A recovery key is automatically generated as part of this setup. If you forget both your password and recovery key, the data will be lost.

FileVault is turned off for the disk "Macintosh HD".

Lockdown Mode — Turn On...

Lockdown Mode is an extreme, optional protection that should only be used if you believe you may be personally targeted by a highly sophisticated cyberattack. Most people are never targeted by attacks of this nature.

When Mac is in Lockdown Mode, it will not function as it typically does. Applications, websites and features will be strictly limited for security, and some experiences will be completely unavailable.

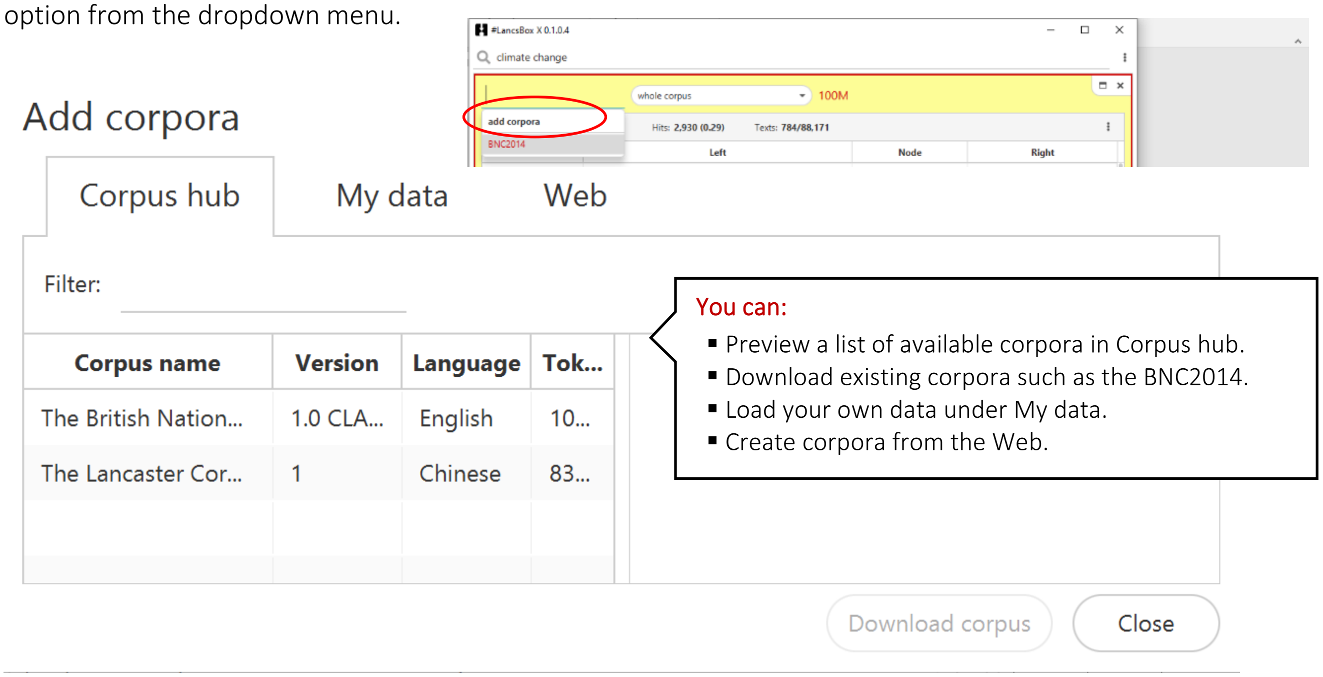
Learn more...

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.



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

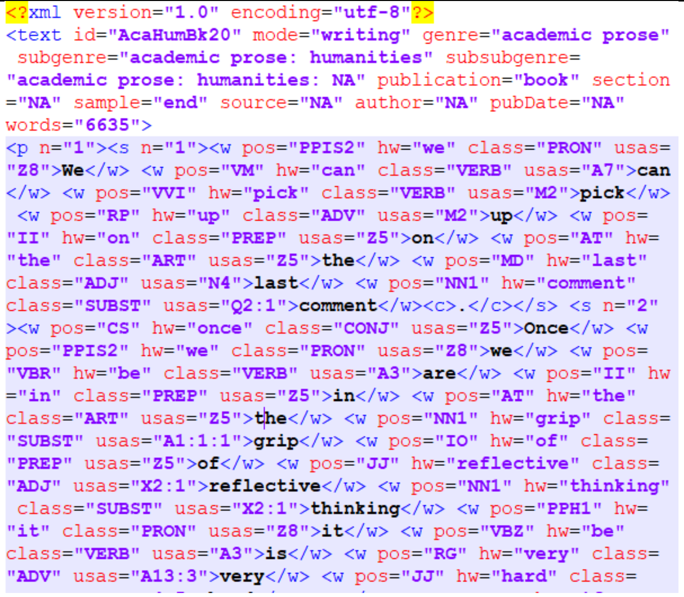
You can:

- 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.

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

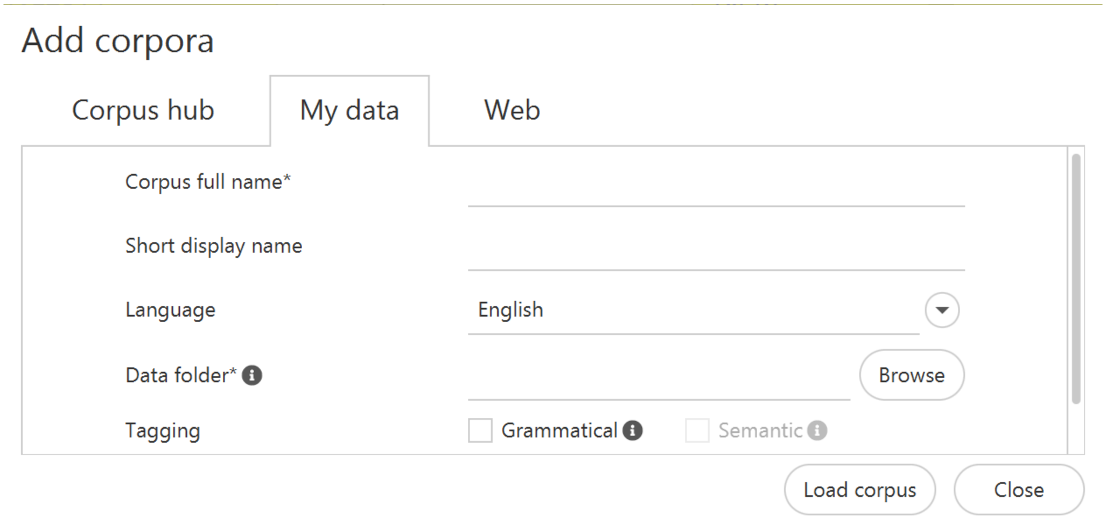
2.2 My data

#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
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.	 <?xml version="1.0" encoding="utf-8"?> <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"> <p n="1"><s n="1"><w pos="PPIS2" hw="we" class="PRON" usas="Z8">We</w> <w pos="VM" hw="can" class="VERB" usas="A7">can</w> <w pos="VVI" hw="pick" class="VERB" usas="M2">pick</w> <w pos="RP" hw="up" class="ADV" usas="M2">up</w> <w pos="II" hw="on" class="PREP" usas="Z5">on</w> <w pos="AT" hw="the" class="ART" usas="Z5">the</w> <w pos="MD" hw="last" class="ADJ" usas="N4">last</w> <w pos="NN1" hw="comment" class="SUBST" usas="Q2:1">comment</w><c>.</c></s> <s n="2"><w pos="CS" hw="once" class="CONJ" usas="Z5">Once</w> <w pos="PPIS2" hw="we" class="PRON" usas="Z8">we</w> <w pos="VBR" hw="be" class="VERB" usas="A3">are</w> <w pos="II" hw="in" class="PREP" usas="Z5">in</w> <w pos="AT" hw="the" class="ART" usas="Z5">the</w> <w pos="NN1" hw="grip" class="SUBST" usas="A1:1:1">grip</w> <w pos="IO" hw="of" class="PREP" usas="Z5">of</w> <w pos="JJ" hw="reflective" class="ADJ" usas="X2:1">reflective</w> <w pos="NN1" hw="thinking" class="SUBST" usas="X2:1">thinking</w> <w pos="PPH1" hw="it" class="PRON" usas="Z8">it</w> <w pos="VBZ" hw="be" class="VERB" usas="A3">is</w> <w pos="RG" hw="very" class="ADV" usas="A13:3">very</w> <w pos="JJ" hw="hard" class=

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 full name*

Short display name

Language: English

Data folder*

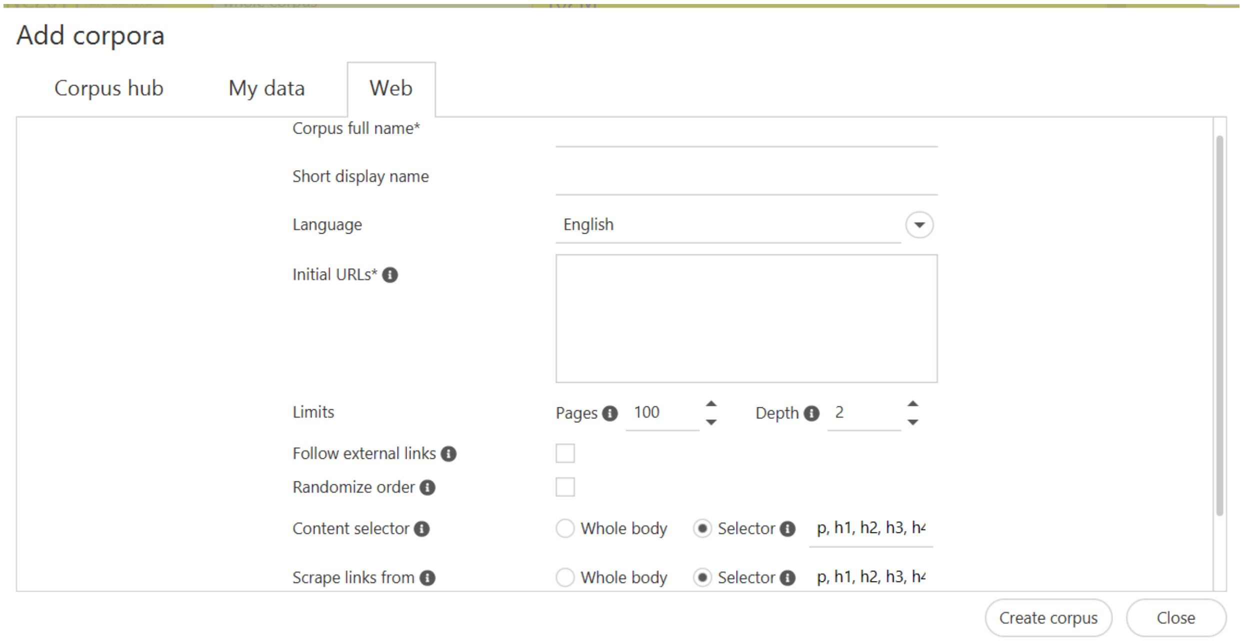
Tagging: □ Grammatical □ Semantic

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 full name*

Short display name

Language: English

Initial URLs*

Limits: Pages 100 Depth 2

Follow external links □

Randomize order □

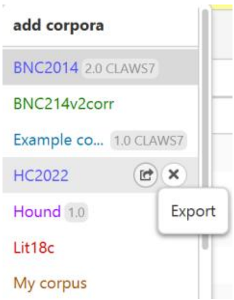
Content selector ○ Whole body ● Selector p, h1, h2, h3, h4

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

2.4 Exporting corpora

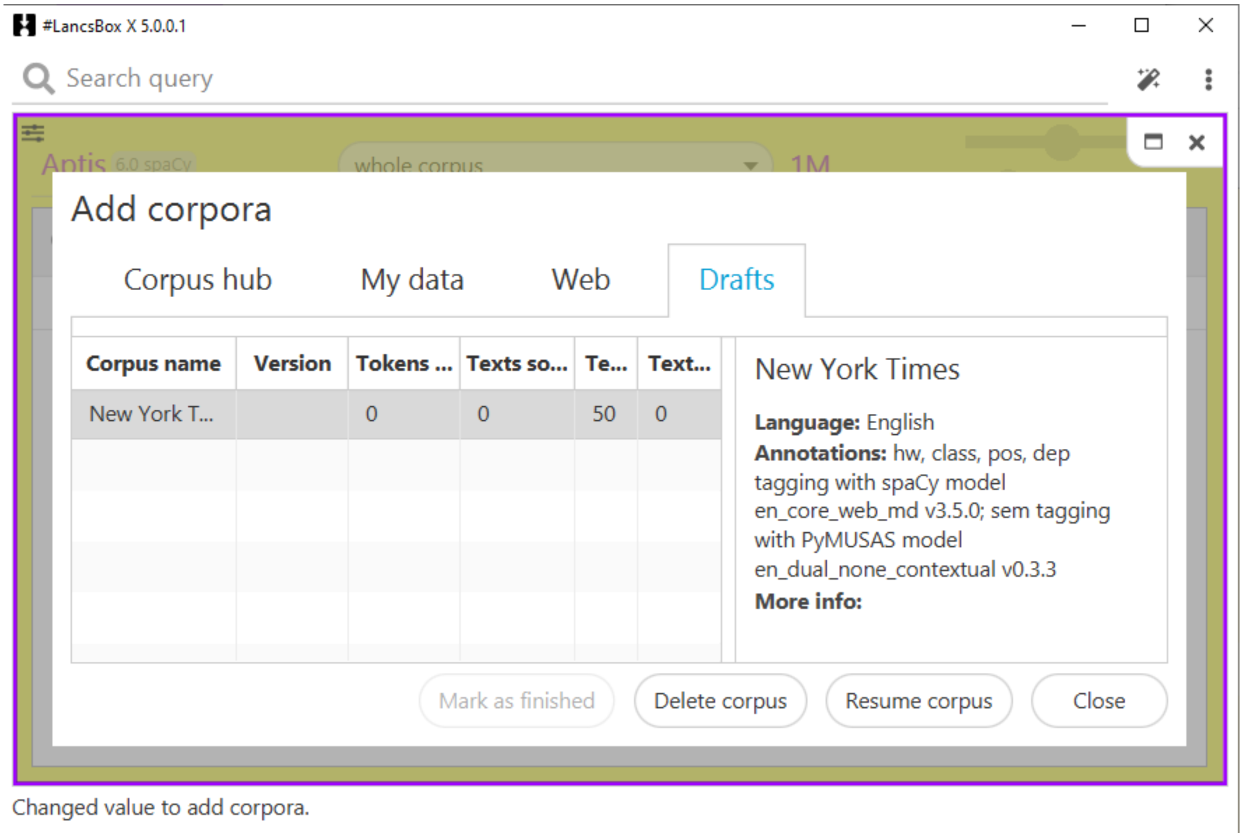
#LancsBox allows you to export corpora in XML. This functionality is available for corpora with unrestricted access.

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



2.5 Draft corpora

#LancsBox allows you to pause corpus processing and return to it later; corpora, which are being processed (and optionally tagged) are also backed up at regular points, which allows returning to the last saved point should something go wrong with the process. Incomplete corpora are available under 'Drafts'



Corpus name	Version	Tokens ...	Texts so...	Te...	Text...
New York T...		0	0	50	0

New York Times

Language: English
Annotations: hw, class, pos, dep tagging with spaCy model en_core_web_md v3.5.0; sem tagging with PyMUSAS model en_dual_none_contextual v0.3.3
More info:

Changed value to add corpora.

To continue processing a corpus, select the appropriate corpus from the list and click on 'Resume corpus'.

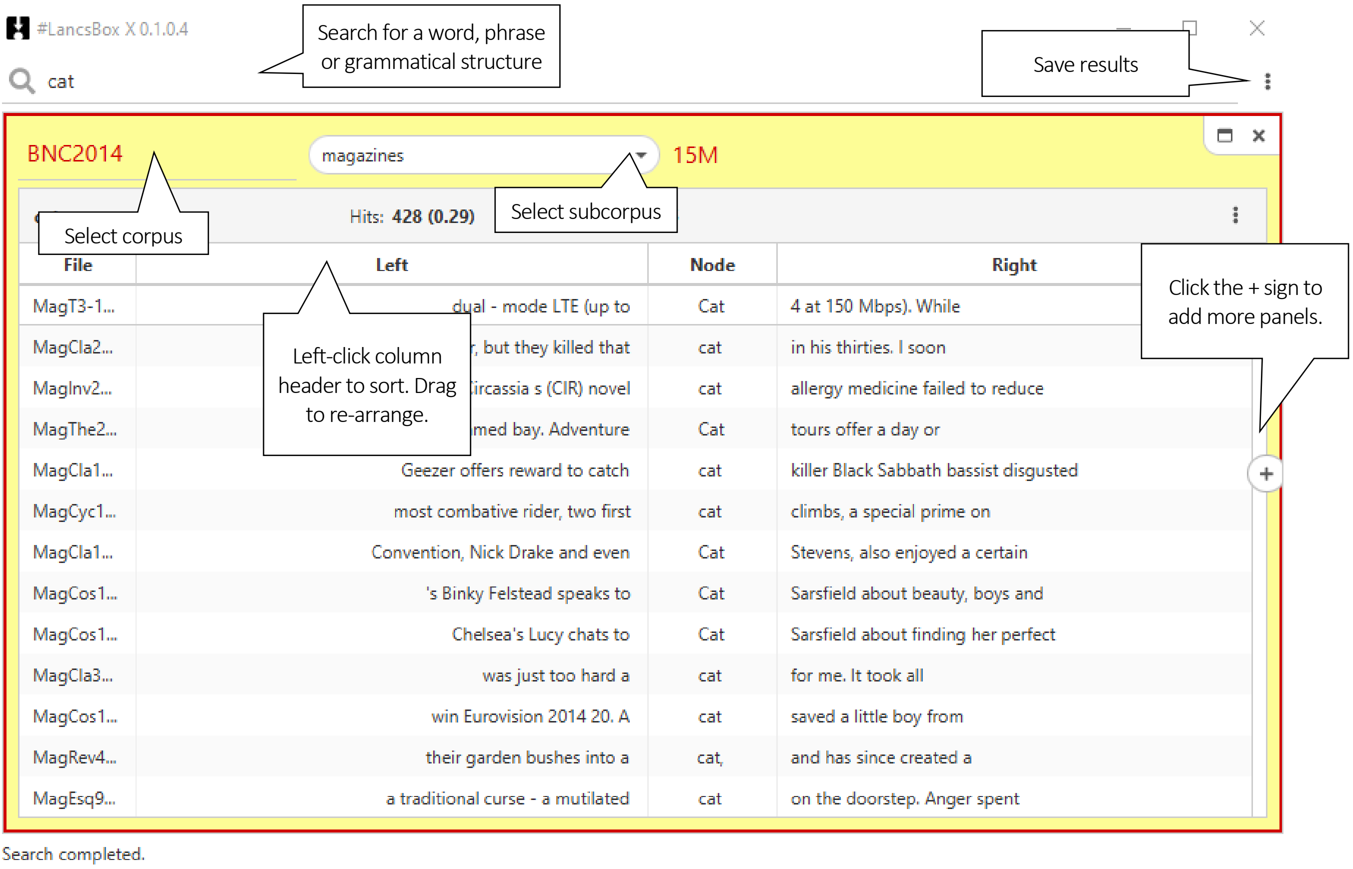
3 KWIC tool (Key Word In Context)

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.



#LancsBox X 0.1.0.4

cat

Search for a word, phrase or grammatical structure

Save results

BNC2014

magazines

15M

Select corpus

Hits: 428 (0.29)

Select subcorpus

Left-click column header to sort. Drag to re-arrange.

Click the + sign to add more panels.

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

Search completed.

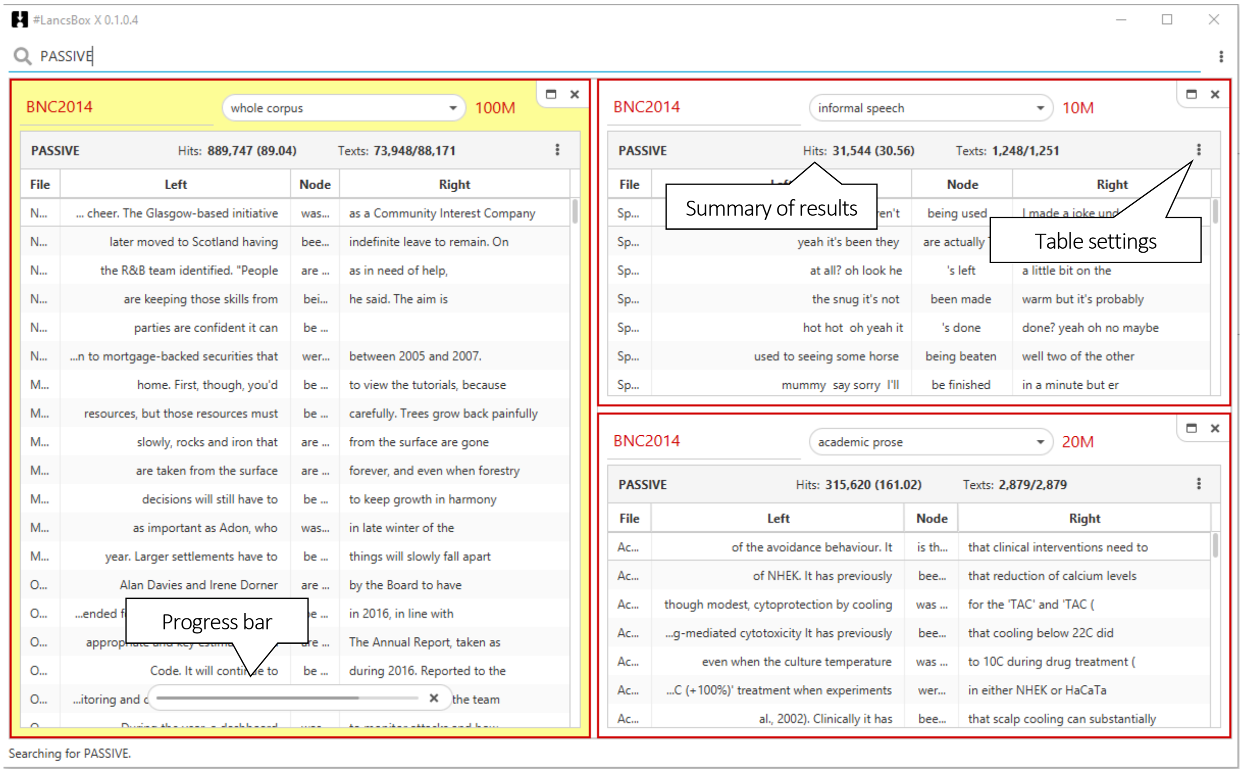
Click a row in a table to select it. Hold the Ctrl or Cmd key while clicking to select multiple rows. Selected rows can be copied with the Ctrl+C / Cmd+C keyboard shortcut or by right clicking the table and selecting the “Copy” option.

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.



PASSIVE

BNC2014 whole corpus 100M

PASSIVE Hits: 889,747 (89.04) Texts: 73,948/88,171

File	Left	Node	Right
N...	... cheer. The Glasgow-based initiative	was...	as a Community Interest Company
N...	later moved to Scotland having	bee...	indefinite leave to remain. On
N...	the R&B team identified. "People	are ...	as in need of help,
N...	are keeping those skills from	bei...	he said. The aim is
N...	parties are confident it can	be ...	
N...	...n to mortgage-backed securities that	wer...	between 2005 and 2007.
M...	home. First, though, you'd	be ...	to view the tutorials, because
M...	resources, but those resources must	be ...	carefully. Trees grow back painfully
M...	slowly, rocks and iron that	are ...	from the surface are gone
M...	are taken from the surface	are ...	forever, and even when forestry
M...	decisions will still have to	be ...	to keep growth in harmony
M...	as important as Adon, who	was...	in late winter of the
M...	year. Larger settlements have to	be ...	things will slowly fall apart
O...	Alan Davies and Irene Dorner	are ...	by the Board to have
O...	...ended f	e ...	in 2016, in line with
O...	appropriate and key estima	are ...	The Annual Report, taken as
O...	Code. It will continue to	be ...	during 2016. Reported to the
O...	...itoring and c		the team
O...	During the year, a dashboard	was	to monitor attacks and how

Progress bar

Searching for PASSIVE.

BNC2014 informal speech 10M

PASSIVE Hits: 31,544 (30.56) Texts: 1,248/1,251

File	Left	Node	Right
Sp...	...en't	being used	I made a joke und
Sp...	yeah it's been they	are actually	
Sp...	at all? oh look he	's left	a little bit on the
Sp...	the snug it's not	been made	warm but it's probably
Sp...	hot hot oh yeah it	's done	done? yeah oh no maybe
Sp...	used to seeing some horse	being beaten	well two of the other
Sp...	mummy say sorry I'll	be finished	in a minute but er

Summary of results

Table settings

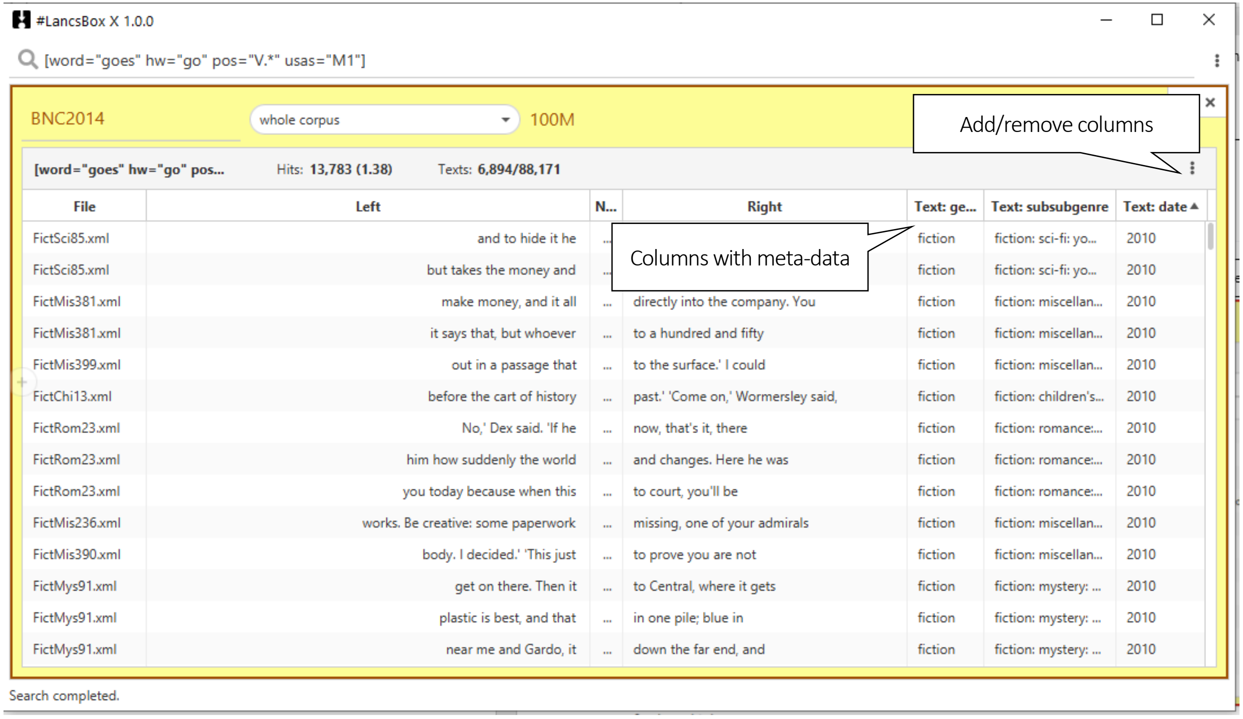
BNC2014 academic prose 20M

PASSIVE Hits: 315,620 (161.02) Texts: 2,879/2,879

File	Left	Node	Right
Ac...	of the avoidance behaviour. It	is th...	that clinical interventions need to
Ac...	of NHEK. It has previously	bee...	that reduction of calcium levels
Ac...	though modest, cytoprotection by cooling	was ...	for the 'TAC' and 'TAC (
Ac...	...g-mediated cytotoxicity It has previously	bee...	that cooling below 22C did
Ac...	even when the culture temperature	was ...	to 10C during drug treatment (
Ac...	...C (+100%)' treatment when experiments	wer...	in either NHEK or HaCaTa
Ac...	al., 2002). Clinically it has	bee...	that scalp cooling can substantially

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.



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

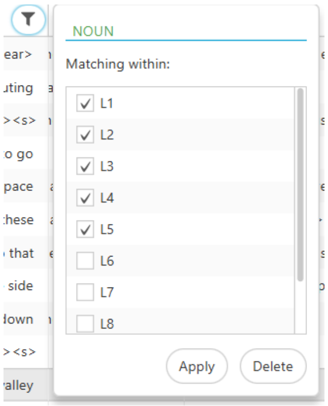
[word="goes" hw="go" pos... Hits: 13,783 (1.38) Texts: 6,894/88,171

File	Left	N...	Right	Text: ge...	Text: subsubgenre	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,' Wormersley 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

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 ▼ .

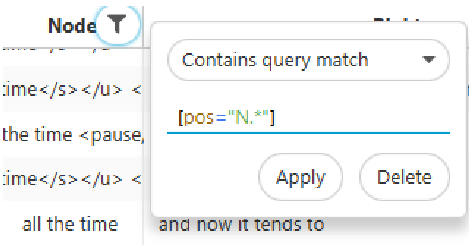
Linguistic data can be filtered using the complete linguistic search functionality. For the left and the right context, choose the position(s) where the required linguistic feature should occur.



NOUN

Matching within:

☑ L1
☑ L2
☑ L3
☑ L4
☑ L5
□ L6
□ L7
□ L8

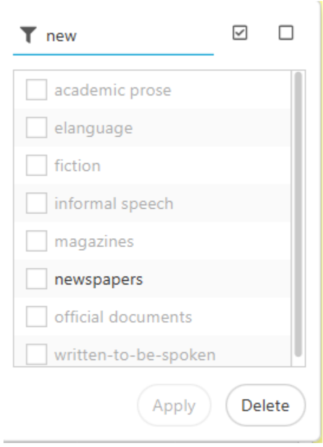


Contains query match

[pos="N.*"]

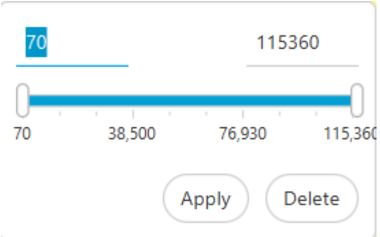
Metalinguistic data can be filtered according to three data types: i) categories, ii) numbers and iii) dates.

Categories



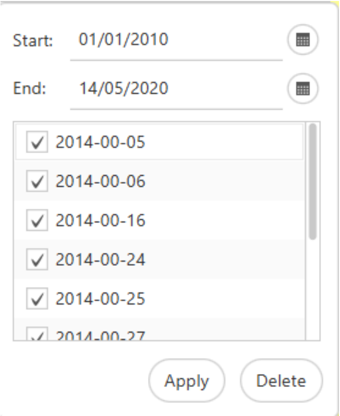
Select required categories by ticking the check box next to each category or search for categories and press the select all highlighted categories button ☑.

Numbers



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

Dates

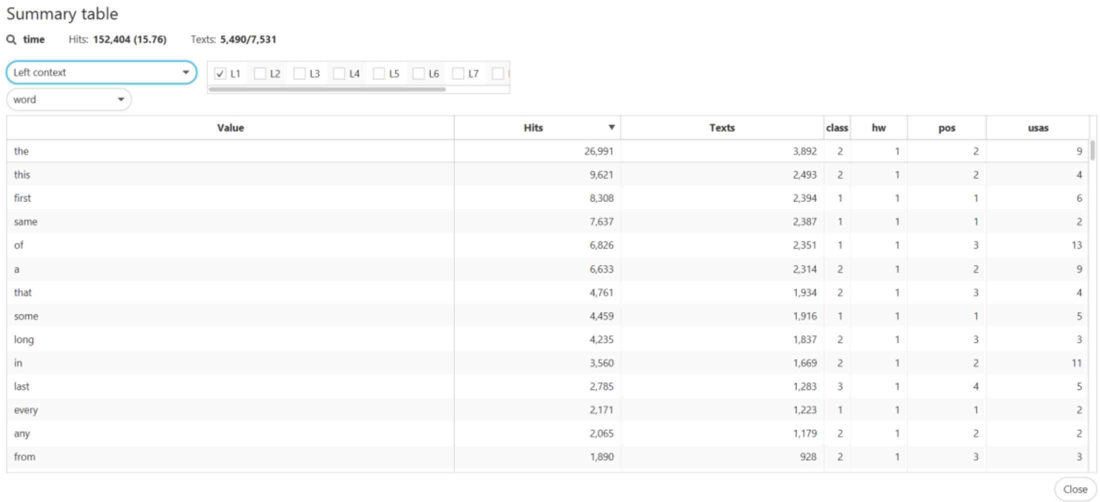


Select a start and end date. Dates that do not follow a valid YYYY-MM-DD pattern are displayed as categories.

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.



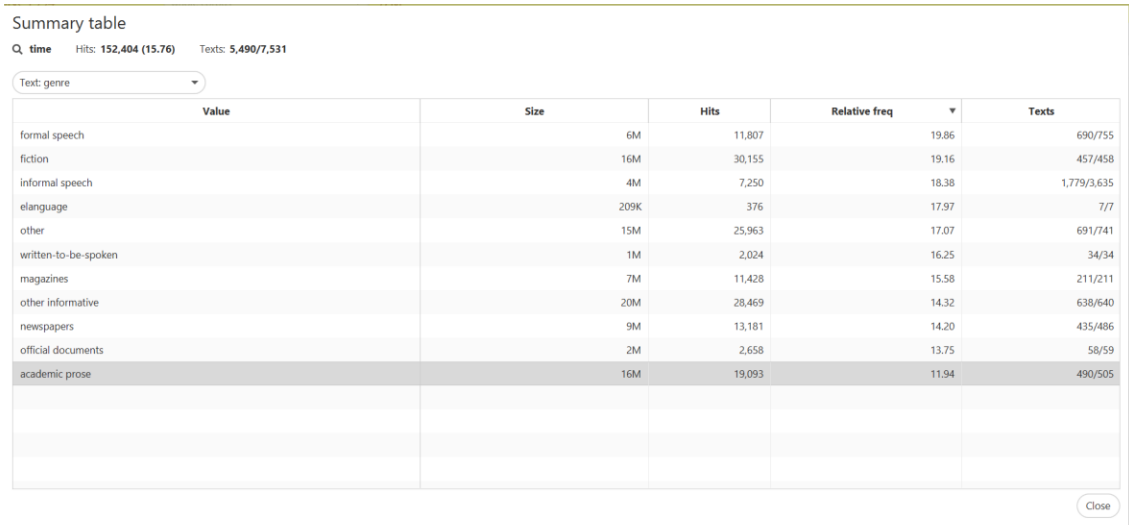
Summary table

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

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

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

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

Text: genre

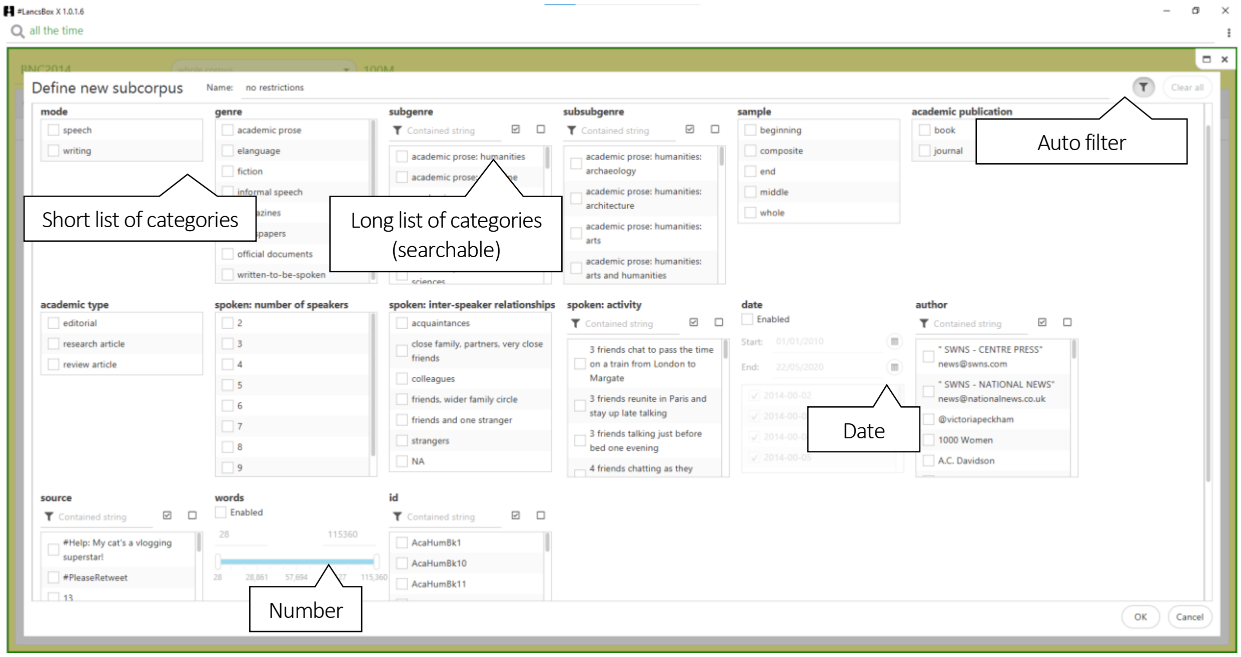
Value	Size	Hits	Relative freq ▾	Texts
formal speech	6M	11,807	19.86	690/755
fiction	16M	30,155	19.16	457/458
informal speech	4M	7,250	18.38	1,779/3,635
elanguage	209K	376	17.97	7/7
other	15M	25,963	17.07	691/741
written-to-be-spoken	1M	2,024	16.25	34/34
magazines	7M	11,428	15.58	211/211
other informative	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

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.



Auto filter

Short list of categories

Long list of categories (searchable)

Date

Number

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.

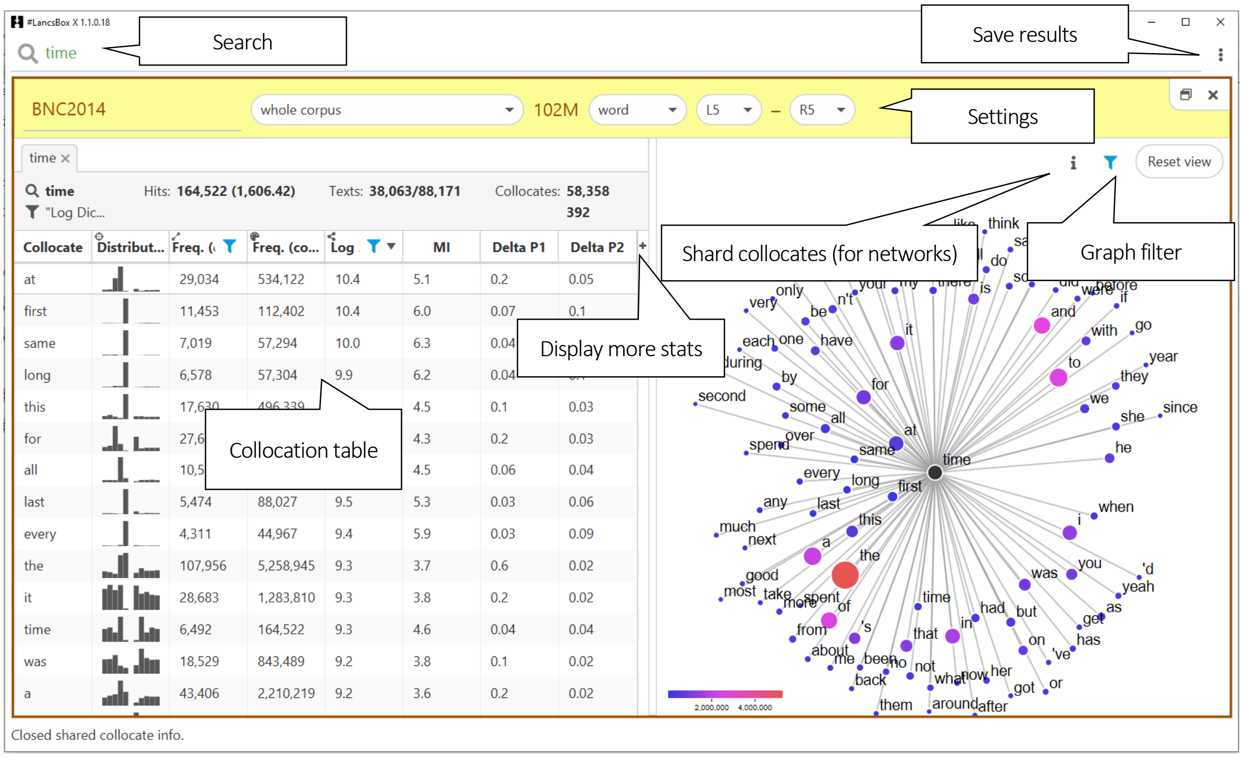
4 GraphColl

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



time   Hits: **164,522 (1,606.42)**   Texts: **38,063/88,171**   Collocates: **58,358**
"Log Dic...   **392**

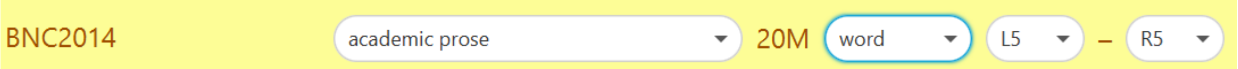
Collocate	Distribut...	Freq. (c	Freq. (co...	Log	MI	Delta P1	Delta P2
at		29,034	534,122	10.4	5.1	0.2	0.05
first		11,453	112,402	10.4	6.0	0.07	0.1
same		7,019	57,294	10.0	6.3	0.04	
long		6,578	57,304	9.9	6.2	0.04	
this		17,630	496,339		4.5	0.1	0.03
for		27,6			4.3	0.2	0.03
all		10,5			4.5	0.06	0.04
last		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

Closed shared collocate info.

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:



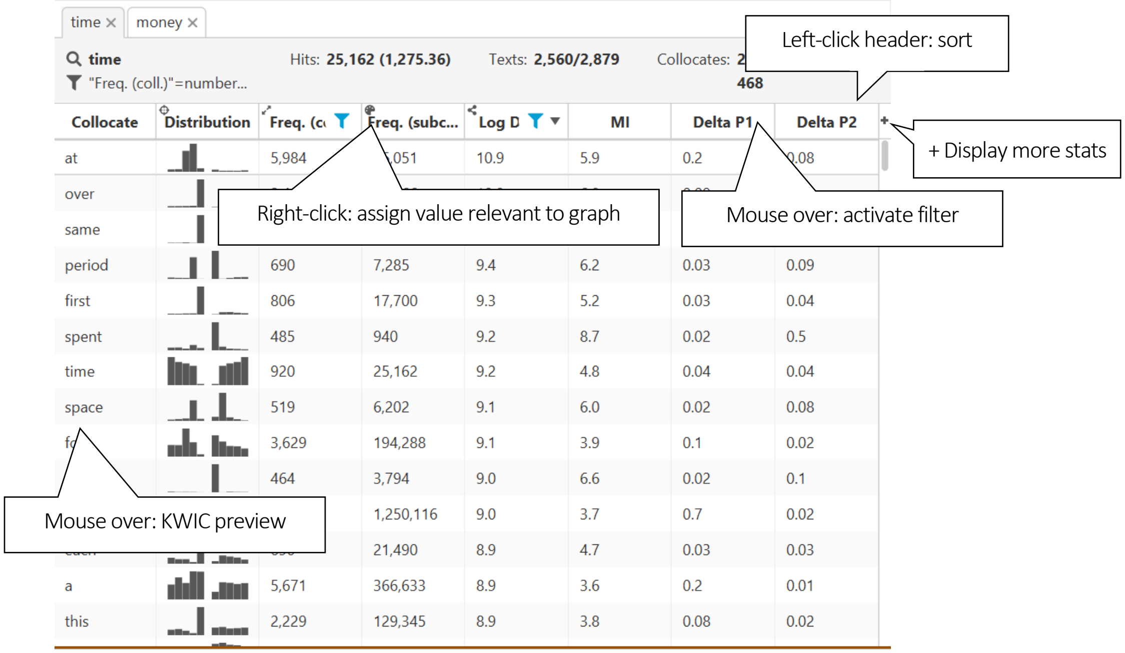
BNC2014 | academic prose | 20M | word | L5 – R5

   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.



time
"Freq. (coll.)"=number...
Hits: 25,162 (1,275.36)   Texts: 2,560/2,879   Collocates: 2[illegible] 468

Collocate	Distribution	Freq. (c	Freq. (subc...	Log D	MI	Delta P1	Delta P2
at		5,984	[illegible]	10.9	5.9	0.2	0.08
over		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
same		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
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
fo[illegible]		3,629	194,288	9.1	3.9	0.1	0.02
[illegible]		464	3,794	9.0	6.6	0.02	0.1
[illegible]	[illegible]	[illegible]	1,250,116	9.0	3.7	0.7	0.02
[illegible]		[illegible]	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

Left-click header: sort

+ Display more stats

Right-click: assign value relevant to graph

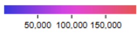
Mouse over: activate filter

Mouse over: KWIC preview

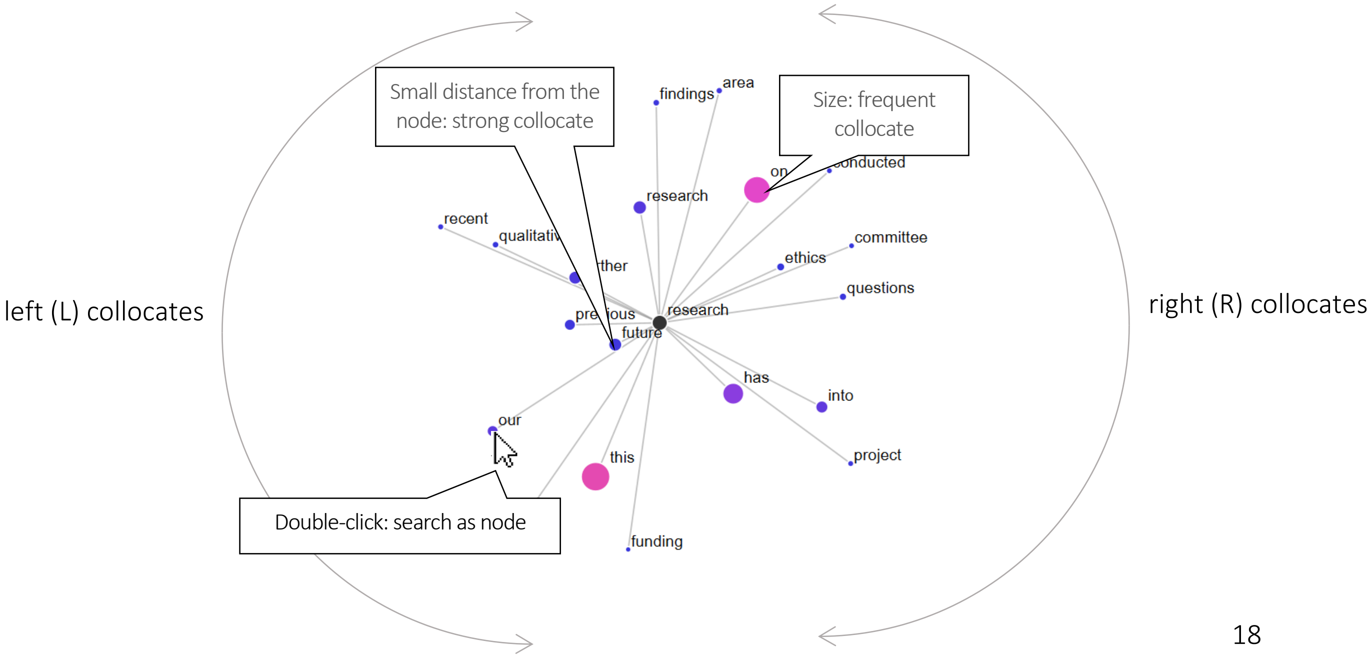
2. The meaning of the individual columns is:
   i) Collocate: shows the collocate in question.
   ii) Distribution: shows a bar chart indicating the textual position of the collocate (e.g. in the L5-R5 span).
   iii) Freq (coll): displays the frequency of the collocation (combination of node + collocate).
   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.


Small distance from the node: strong collocate

Size: frequent collocate

Double-click: search as node

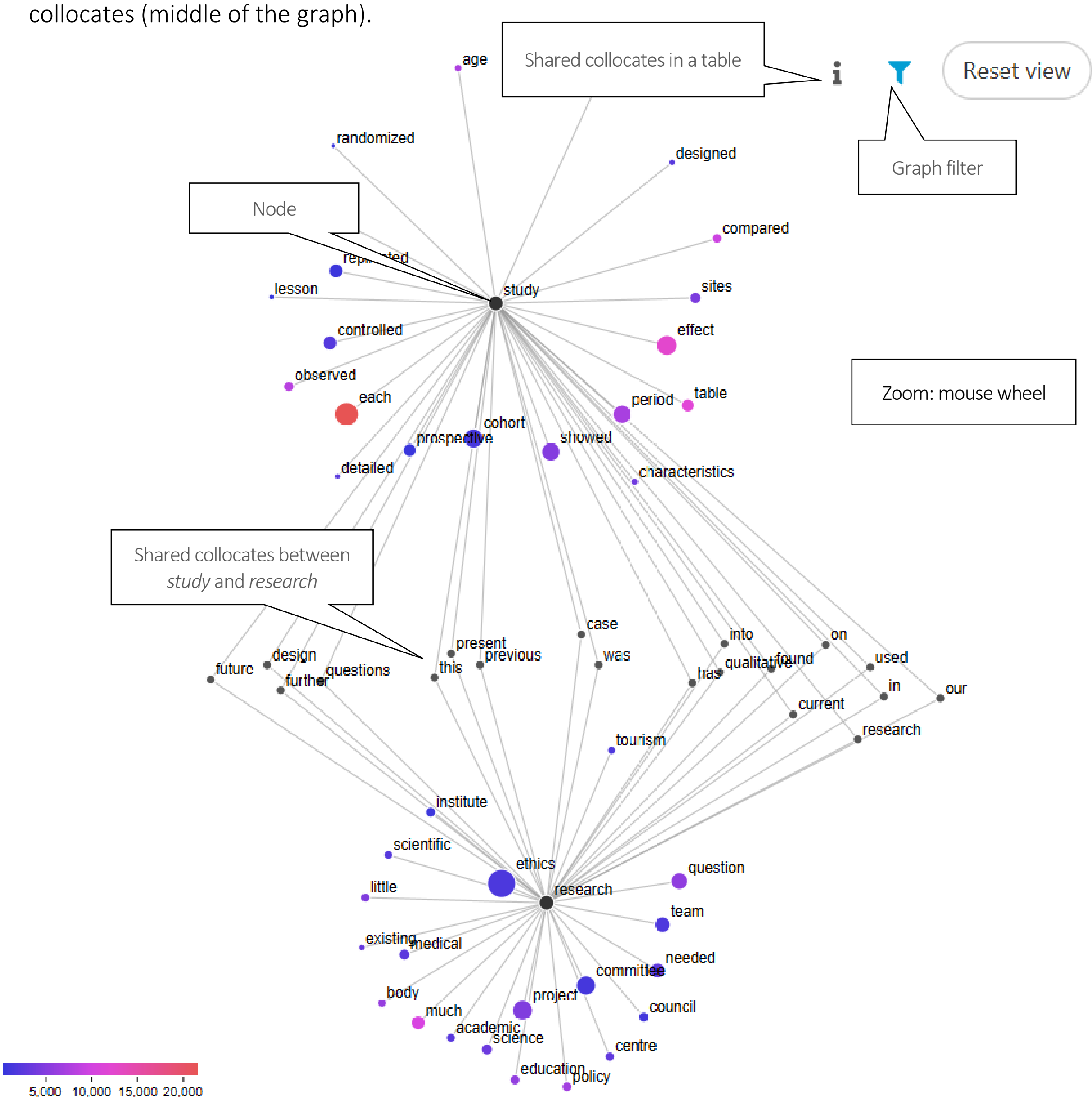
left (L) collocates

right (R) collocates

4.5 Extending graph to a collocation network

A collocation network is an extended collocation graph that shows i) shared collocates and ii) cross-associations between several nodes.

1. To expand a simple collocation graph into a collocation network, either search for more nodes or left-double-click on a collocate in the graph.
2. A collocation network displays nodes with unique collocates (outer rim of the graph) and shared collocates (middle of the graph).


Shared collocates in a table

Graph filter

Reset view

Node

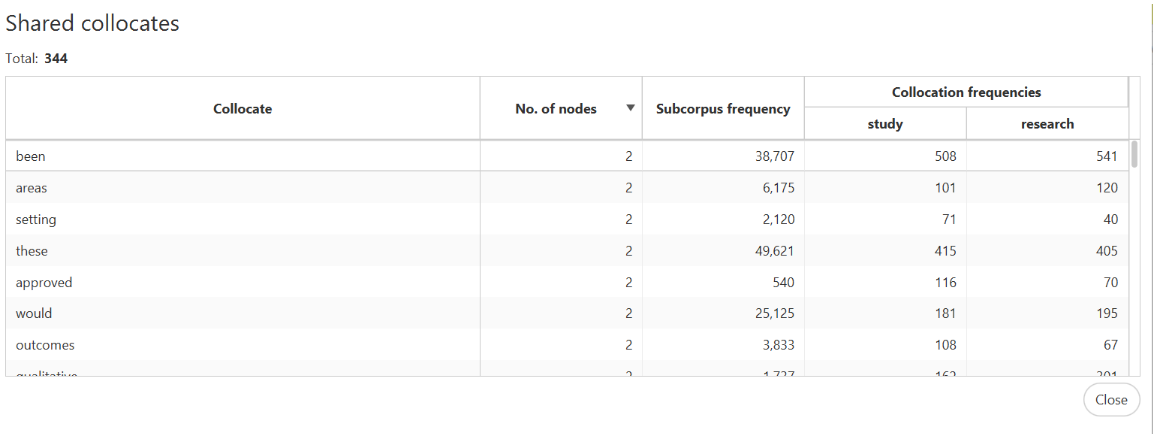
Zoom: mouse wheel

Shared collocates between *study* and *research*

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 ℹ .
2. The list of shard collocates is displayed in a tabular form.



Shared collocates

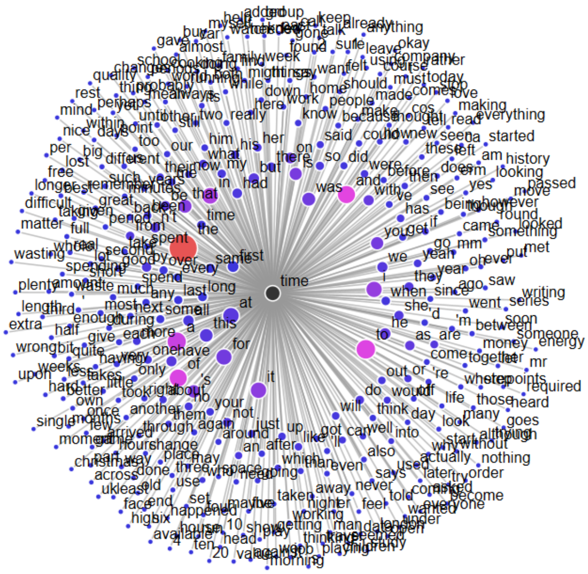
Total: **344**

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
qualitative	2	1,727	162	201

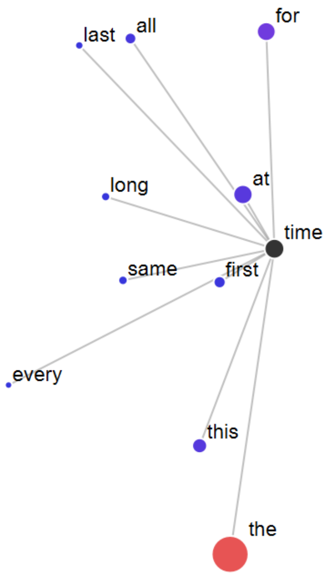
4.7 Problems with graphs: overpopulated graphs

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.

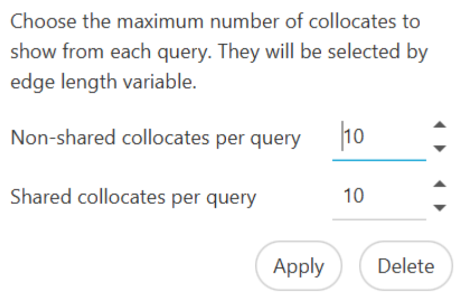
The following figure shows an overpopulated graph on the left and a graph that is more easily interpretable on the right.



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: 10

Shared collocates per query: 10

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 *graph*ical *coll*ocations 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.

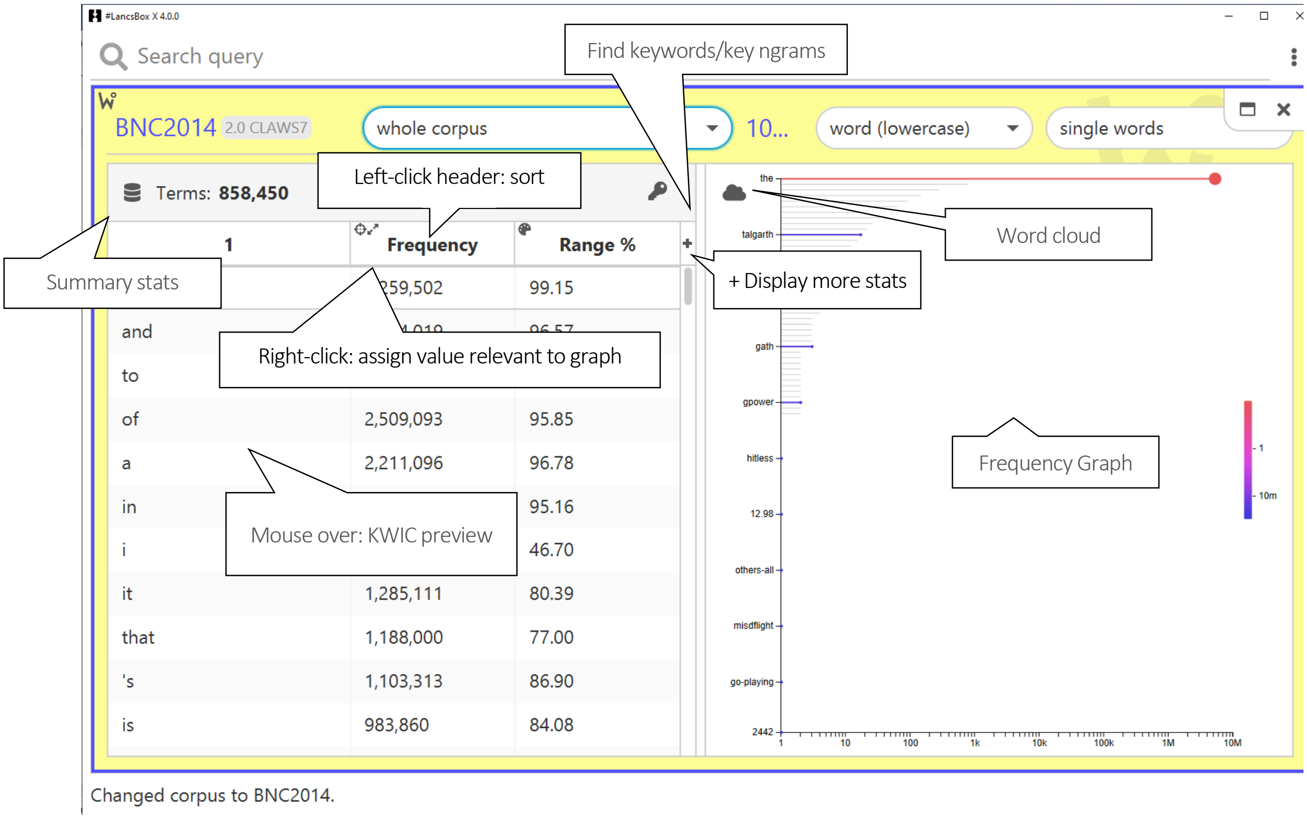
5 Words tool

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



1	Frequency	Range %
	259,502	99.15
and	1,019	96.57
to		
of	2,509,093	95.85
a	2,211,096	96.78
in		95.16
i		46.70
it	1,285,111	80.39
that	1,188,000	77.00
's	1,103,313	86.90
is	983,860	84.08

Changed corpus to BNC2014.

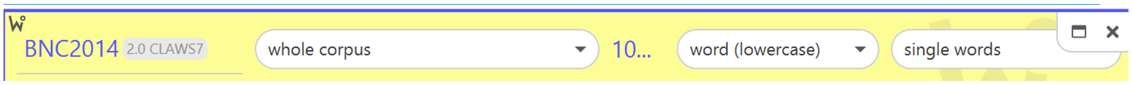
Left: Creating frequency lists, computing dispersion and keywords.

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:

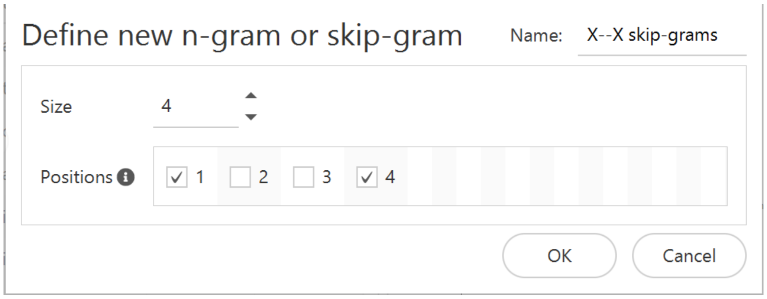


BNC2014 2.0 CLAWS7 | whole corpus | 10... | word (lowercase) | single words

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.



Define new n-gram or skip-gram

Name: X--X skip-grams

Size: 4

Positions: ☑ 1 □ 2 □ 3 ☑ 4

OK | Cancel

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.

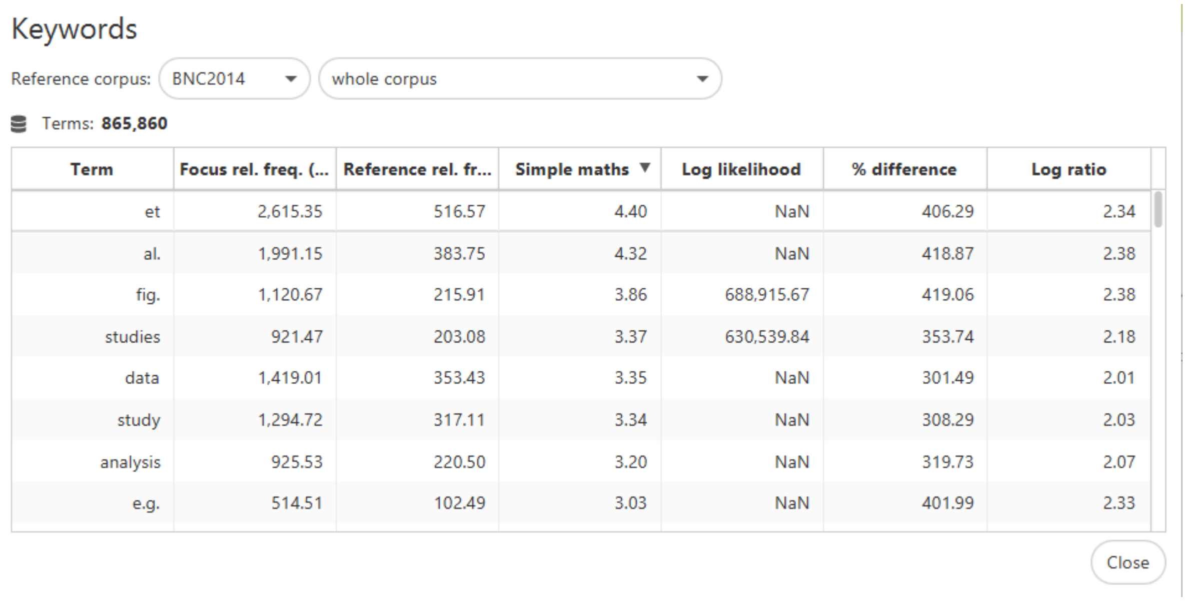
Note: Please note that Frequency lists in #LancsBox X are pre-computed and stored for later use. If you are creating a wordlist for the first time, this might take some time depending on the size of the corpus and complexity of its annotation (number of units used).

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.

3. Sort and/or filter according to your preferred keyword statistics (Simple maths is used by default for sorting).



Keywords

Reference corpus: BNC2014 ▾ whole corpus ▾

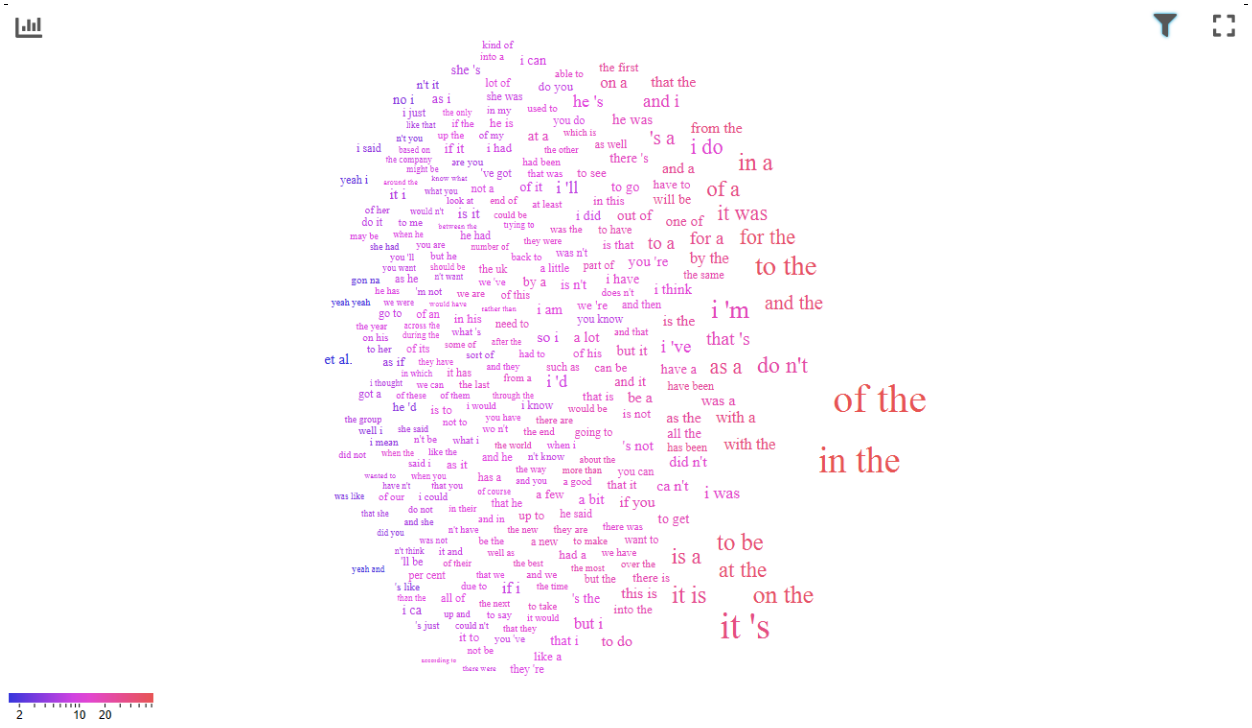
Terms: **865,860**

Term	Focus rel. freq. (...	Reference rel. fr...	Simple maths ▼	Log likelihood	% difference	Log ratio
et	2,615.35	516.57	4.40	NaN	406.29	2.34
al.	1,991.15	383.75	4.32	NaN	418.87	2.38
fig.	1,120.67	215.91	3.86	688,915.67	419.06	2.38
studies	921.47	203.08	3.37	630,539.84	353.74	2.18
data	1,419.01	353.43	3.35	NaN	301.49	2.01
study	1,294.72	317.11	3.34	NaN	308.29	2.03
analysis	925.53	220.50	3.20	NaN	319.73	2.07
e.g.	514.51	102.49	3.03	NaN	401.99	2.33

Close

5.4 Word cloud

The Words module creates word clouds based on words, n-grams, grammatical and semantic structures. Word clouds can be assigned different statistical properties from the table indicated by i) position, ii) font size and iii) colour in the graph.



▶ 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*.

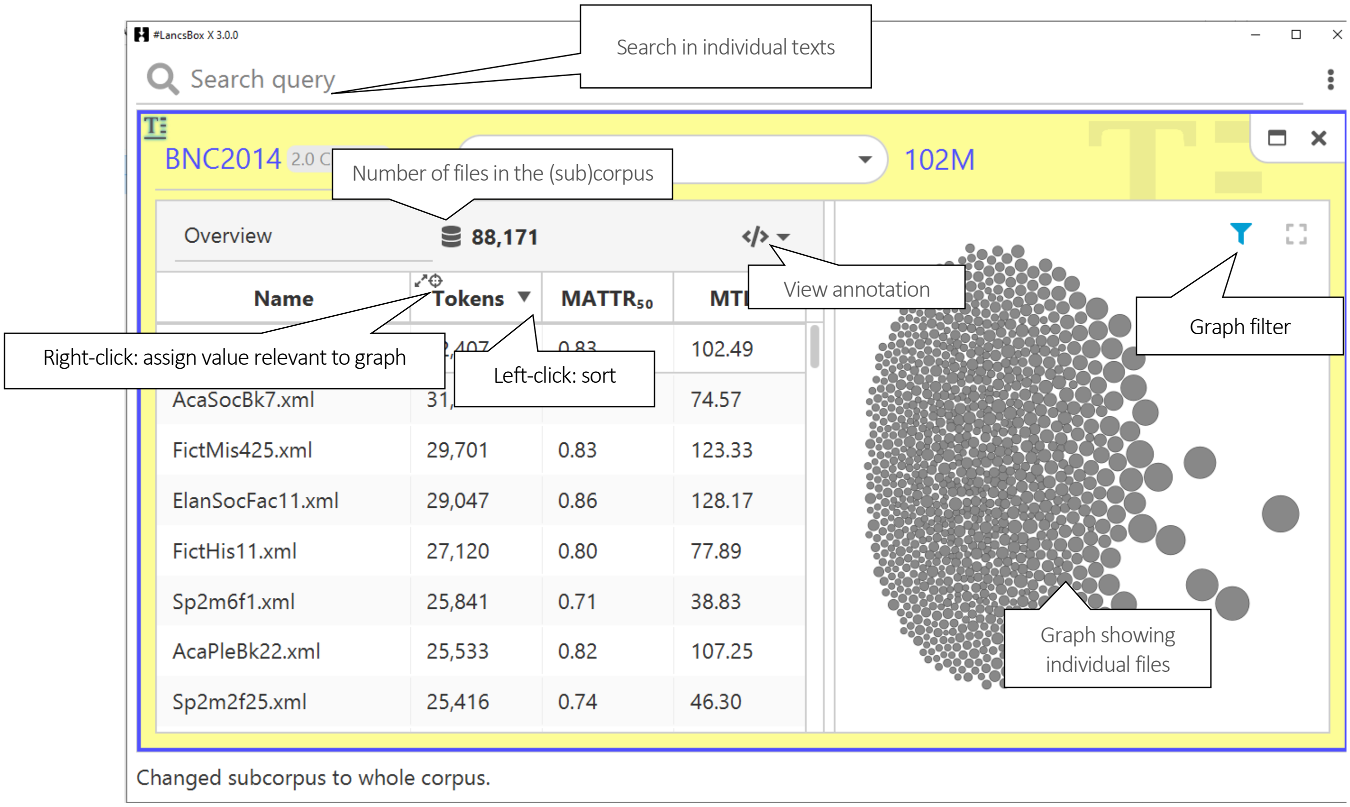
6 Text tool

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



Search in individual texts

Number of files in the (sub)corpus

View annotation

Graph filter

Right-click: assign value relevant to graph

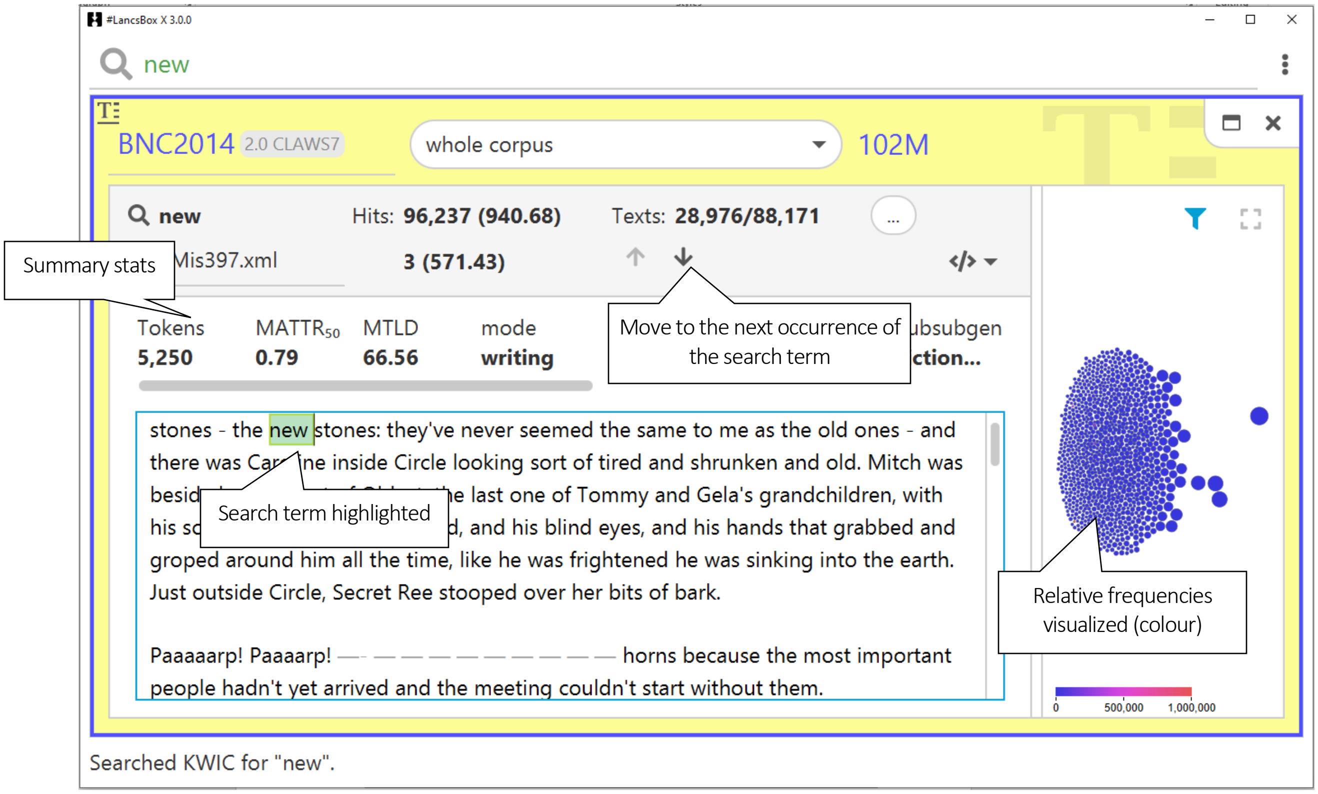
Left-click: sort

Graph showing individual files

Name	Tokens ▼	MATTR$_{50}$	MTL
	2,407	0.83	102.49
AcaSocBk7.xml	31,		74.57
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



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

Mis397.xml 3 (571.43)

Tokens	MATTR50	MTLD	mode	ubsubgen
5,250	0.79	66.56	writing	ction...

stones - the new stones: they've never seemed the same to me as the old ones - and
there was Car ne inside Circle looking sort of tired and shrunken and old. Mitch was
besid e last one of Tommy and Gela's grandchildren, with
his sc d, 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! Paaaarp! — — — — — — — — — — — horns because the most important
people hadn't yet arrived and the meeting couldn't start without them.

Searched KWIC for "new".

7 Wizard

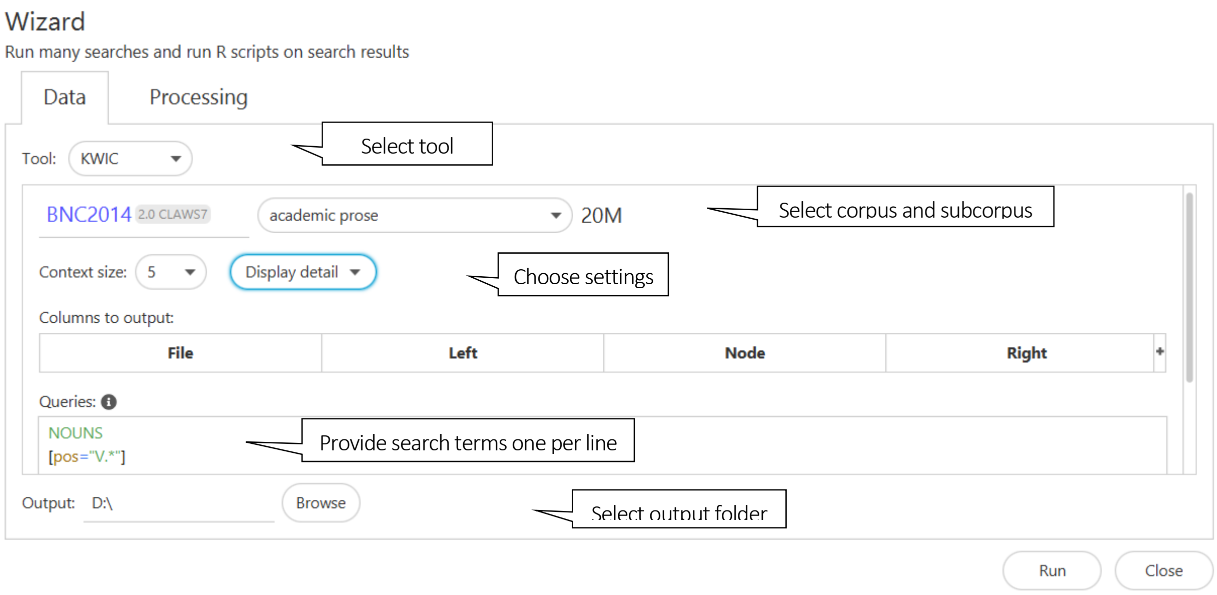
The Wizard tool allows batch searching of corpora and running statistical analyses on the results. Wizard implements the R package, which can be used to run simple and complex statistical analyses inside #LancsBox. To start the Wizard tool, click on the Wizard icon in the top right corner of the search bar.

It can be used, for example, to:

- Search for multiple search terms at once and save the results.
- Search in multiple corpora at once and save the results.
- Search multiple tools (KWIC, GraphColl, Words, Text) at once and search the results.
- Statistically analyse results.
- Visualise results.

7.1 Wizard: An overview

Data tab



Wizard

Run many searches and run R scripts on search results

Tool: KWIC

BNC2014 2.0 CLAWS7 | academic prose | 20M

Context size: 5 | Display detail

Columns to output:

File	Left	Node	Right

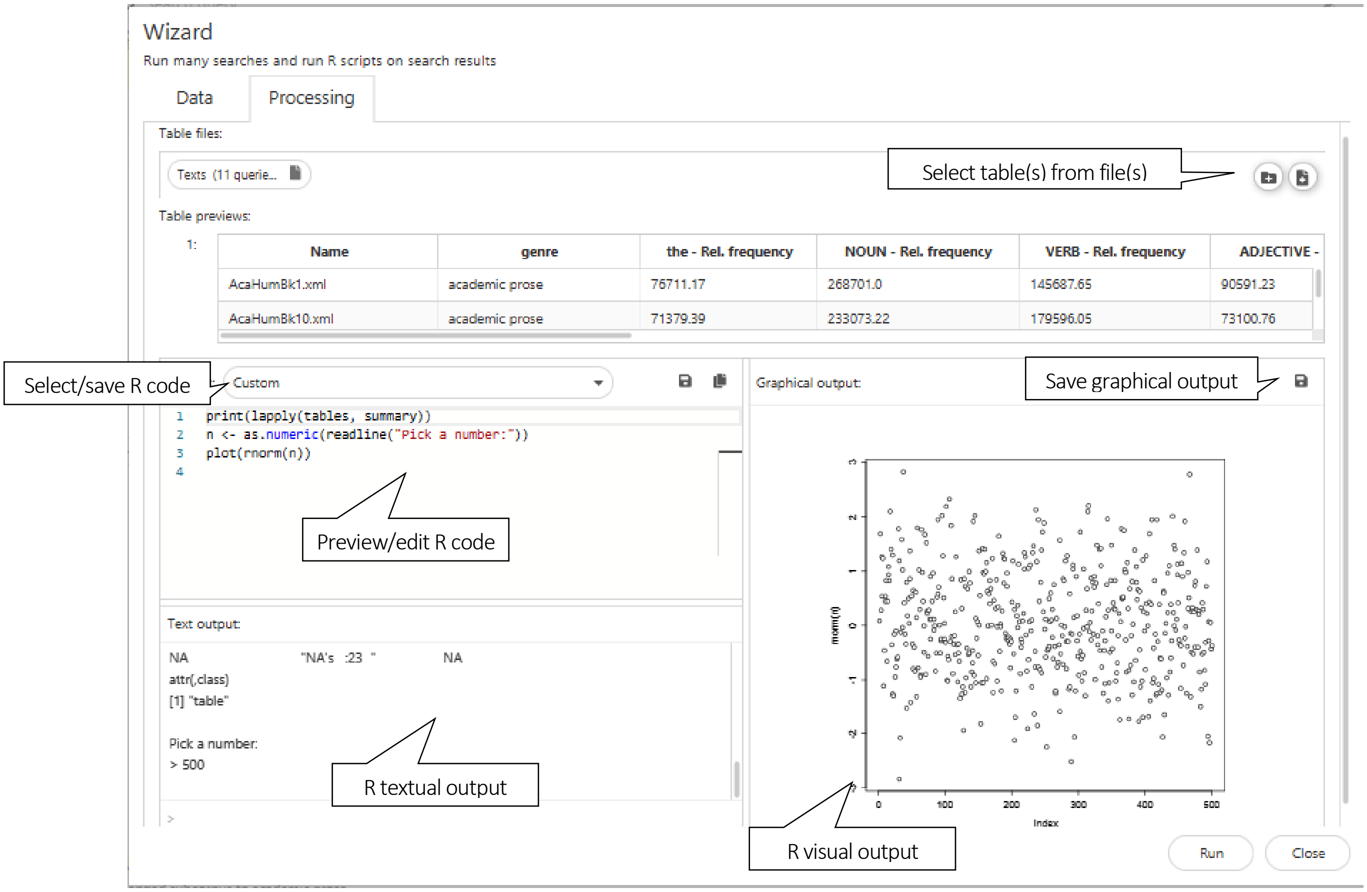
Queries:

```
NOUNS
[pos="V.*"]
```

Output: D:\

All search processes from the Data tab run in the background and are displayed in the bottom right corner of the tool. Progress is indicated by a blue circle; running searches can be cancelled by clicking on the ✕ cancel button into which the icon turns on mouse over.

Processing tab



Name	genre	the - Rel. frequency	NOUN - Rel. frequency	VERB - Rel. frequency	ADJECTIVE -
AcaHumBk1.xml	academic prose	76711.17	268701.0	145687.65	90591.23
AcaHumBk10.xml	academic prose	71379.39	233073.22	179596.05	73100.76

```
print(lapply(tables, summary))
n <- as.numeric(readline("Pick a number:"))
plot(rnorm(n))
```

```
NA                  "NA's  :23  "           NA
attr(,class)
[1] "table"

Pick a number:
> 500
```

Select table(s) from file(s)

Select/save R code

Save graphical output

Preview/edit R code

R textual output

R visual output

7.2 R code

1. To refer to individual tables from previewed Tables, use tables[[1]], tables[[2]], tables[[3]]

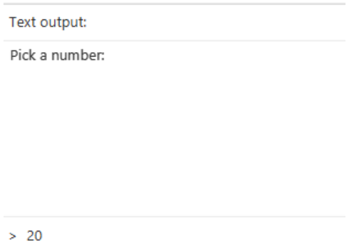
```
#Step 1: get data from Table 1
data <-tables[[1]]
```

2. To print an output, use ‘print’.

```
print(tables[[1]])
```

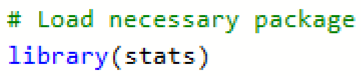
3. To request user input use ‘readline’; provide input and press enter.

```
n <- as.numeric(readline("Pick a number:"))
```



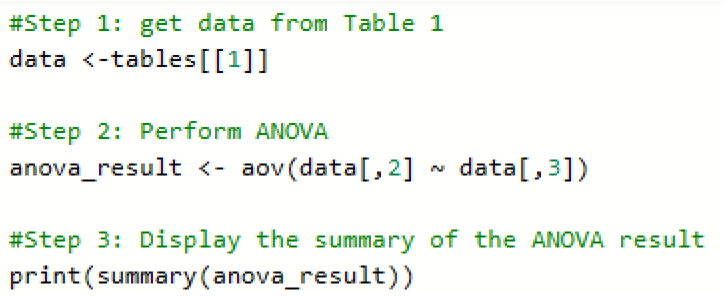
```
Text output:
Pick a number:

> 20
```

4. To load an R library, use ‘library(name_od_library)’; not all libraries are currently supported. For a list of supported libraries (please read the small print on the functionality within those libraries), please see: https://packages.renjin.org/packages. Much of the functionality from individual libraries, if not currently supported, can be taken over by core R functions, which is always available.



```
# Load necessary package
library(stats)
```

5. Here is an example of a complete script performing the ANOVA statistical test.



```
#Step 1: get data from Table 1
data <-tables[[1]]

#Step 2: Perform ANOVA
anova_result <- aov(data[,2] ~ data[,3])

#Step 3: Display the summary of the ANOVA result
print(summary(anova_result))
```

8 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* [*new, news, newly, newspaper…*]
	any word [with space]	new *[*new car, New York, new ideas…*]

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 backlash 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"/>)
```

9 spaCy POS tagset: English

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

10 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

11 CLAWS tagset (C7)

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

APPGE	possessive pronoun, pre-nominal (e.g. my, your, our)
AT	article (e.g. the, no)
AT1	singular article (e.g. a, an, every)
BCL	before-clause marker (e.g. in order (that),in order (to))
CC	coordinating conjunction (e.g. and, or)
CCB	adversative coordinating conjunction (but)
CS	subordinating conjunction (e.g. if, because, unless, so, for)
CSA	as (as conjunction)
CSN	than (as conjunction)
CST	that (as conjunction)
CSW	whether (as conjunction)
DA	after-determiner or post-determiner capable of pronominal function (e.g. such, former, same)
DA1	singular after-determiner (e.g. little, much)
DA2	plural after-determiner (e.g. few, several, many)
DAR	comparative after-determiner (e.g. more, less, fewer)
DAT	superlative after-determiner (e.g. most, least, fewest)
DB	before determiner or pre-determiner capable of pronominal function (all, half)
DB2	plural before-determiner (both)
DD	determiner (capable of pronominal function) (e.g any, some)
DD1	singular determiner (e.g. this, that, another)
DD2	plural determiner (these,those)
DDQ	wh-determiner (which, what)
DDQGE	wh-determiner, genitive (whose)
DDQV	wh-ever determiner, (whichever, whatever)
EX	existential there
FO	formula
FU	unclassified word
FW	foreign word
GE	germanic genitive marker - (' or's)
IF	for (as preposition)
II	general preposition
IO	of (as preposition)
IW	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)
NNU	unit of measurement, neutral for number (e.g. in, cc)
NNU1	singular unit of measurement (e.g. inch, centimetre)
NNU2	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

VBN	been
VBR	are
VBZ	is
VD0	do, base form (finite)
VDD	did
VDG	doing
VDI	do, infinitive (I may do... To do...)
VDN	done
VDZ	does
VH0	have, base form (finite)
VHD	had (past tense)
VHG	having
VHI	have, infinitive
VHN	had (past participle)
VHZ	has
VM	modal auxiliary (can, will, would, etc.)
VMK	modal catenative (ought, used)
VV0	base form of lexical verb (e.g. give, work)
VVD	past tense of lexical verb (e.g. gave, worked)
VVG	-ing participle of lexical verb (e.g. giving, working)
VVGK	-ing participle catenative (going in be going to)
VVI	infinitive (e.g. to give... It will work...)
VVN	past participle of lexical verb (e.g. given, worked)
VVNK	past participle catenative (e.g. bound in be bound to)
VVZ	-s form of lexical verb (e.g. gives, works)
XX	not, n't
ZZ1	singular letter of the alphabet (e.g. A,b)
ZZ2	plural letter of the alphabet (e.g. A's, b's)

12 USAS semantic tagset

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

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

I3.1	Work and employment: Generally
I3.2	Work and employmeny: Professionalism
I4	Industry
K1	Entertainment generally
K2	Music and related activities
K3	Recorded sound etc.
K4	Drama, the theatre and showbusiness
K5	Sports and games generally
K5.1	Sports
K5.2	Games
K6	Childrens games and toys
L1	Life and living things
L2	Living creatures generally
L3	Plants
M1	Moving, coming and going
M2	Putting, taking, pulling, pushing, transporting &c.
M3	Vehicles and transport on land
M4	Shipping, swimming etc.
M5	Aircraft and flying
M6	Location and direction
M7	Places
M8	Remaining/stationary
N1	Numbers
N2	Mathematics
N3	Measurement
N3.1	Measurement: General
N3.2	Measurement: Size
N3.3	Measurement: Distance
N3.4	Measurement: Volume
N3.5	Measurement: Weight
N3.6	Measurement: Area
N3.7	Measurement: Length & height
N3.8	Measurement: Speed
N4	Linear order
N5	Quantities
N5.1	Entirety; maximum
N5.2	Exceeding; waste
N6	Frequency etc.
O1	Substances and materials generally
O1.1	Substances and materials generally: Solid
O1.2	Substances and materials generally: Liquid
O1.3	Substances and materials generally: Gas
O2	Objects generally
O3	Electricity and electrical equipment
O4	Physical attributes
O4.1	General appearance and physical properties
O4.2	Judgement of appearance (pretty etc.)
O4.3	Colour and colour patterns
O4.4	Shape
O4.5	Texture
O4.6	Temperature
P1	Education in general
Q1	LINGUISTIC ACTIONS, STATES AND PROCESSES; COMMUNICATION
Q1.1	LINGUISTIC ACTIONS, STATES AND PROCESSES; COMMUNICATION
Q1.2	Paper documents and writing
Q1.3	Telecommunications
Q2	Speech acts
Q2.1	Speech etc:- Communicative
Q2.2	Speech acts
Q3	Language, speech and grammar
Q4	The Media
Q4.1	The Media:- Books
Q4.2	The Media:- Newspapers etc.
Q4.3	The Media:- TV, Radio and Cinema
S1	SOCIAL ACTIONS, STATES AND PROCESSES
S1.1	SOCIAL ACTIONS, STATES AND PROCESSES
S1.1.1	SOCIAL ACTIONS, STATES AND PROCESSES
S1.1.2	Reciprocity
S1.1.3	Participation
S1.1.4	Deserve etc.
S1.2	Personality traits
S1.2.1	Approachability and Friendliness
S1.2.2	Avarice
S1.2.3	Egoism
S1.2.4	Politeness
S1.2.5	Toughness; strong/weak
S1.2.6	Sensible
S2	People
S2.1	People:- Female
S2.2	People:- Male
S3	Relationship
S3.1	Relationship: General
S3.2	Relationship: Intimate/sexual
S4	Kin
S5	Groups and affiliation
S6	Obligation and necessity
S7	Power relationship
S7.1	Power, organizing
S7.2	Respect
S7.3	Competition
S7.4	Permission
S8	Helping/hindering
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
T1.2	Time: Momentary
T1.3	Time: Period
T2	Time: Beginning and ending
T3	Time: Old, new and young; age
T4	Time: Early/late
W1	The universe
W2	Light
W3	Geographical terms
W4	Weather
W5	Green issues
X1	PSYCHOLOGICAL ACTIONS, STATES AND PROCESSES
X2	Mental actions and processes
X2.1	Thought, belief
X2.2	Knowledge
X2.3	Learn
X2.4	Investigate, examine, test, search
X2.5	Understand
X2.6	Expect
X3	Sensory
X3.1	Sensory:- Taste
X3.2	Sensory:- Sound
X3.3	Sensory:- Touch
X3.4	Sensory:- Sight
X3.5	Sensory:- Smell
X4	Mental object
X4.1	Mental object:- Conceptual object
X4.2	Mental object:- Means, method
X5	Attention
X5.1	Attention
X5.2	Interest/boredom/excited/energetic
X6	Deciding
X7	Wanting; planning; choosing
X8	Trying
X9	Ability
X9.1	Ability:- Ability, intelligence
X9.2	Ability:- Success and failure
Y1	Science and technology in general
Y2	Information technology and computing
Z0	Unmatched proper noun
Z1	Personal names
Z2	Geographical names
Z3	Other proper names
Z4	Discourse Bin
Z5	Grammatical bin
Z6	Negative
Z7	If
Z8	Pronouns etc.
Z9	Trash can
Z99	Unmatched

13 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|doading|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|fesnying|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|holiness|horde|host|house|hover|huddle|hunt|hurtle|husk|illusion|implausibility|index|infestation|intrusion|invention|kaleidoscope|kendle|kennel|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|paddling|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|prettying|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|siege|singular|sizzle|skein|skirl|skulk|slate|sleuth|slew|slither|sloth|smack|snarl|snatch|sneak|sord|sounder|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|suite|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|zeal|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.*"]

<table>
<tr><td>DOWNTONER</td><td>[hw="almost|barely|hardly|merely|mildly|nearly|only|partially|partly|practically|scarcely|slightly|somewhat"]</td></tr>
<tr><td>EXISTENTIAL_THERE</td><td>[pos="EX"]</td></tr>
<tr><td>GERUND</td><td>[hw="(?!(.*thing|evening|morning|viking)).{2,}ing" pos="NN[12]"]</td></tr>
<tr><td>HAVE</td><td>[pos="VH.*"]</td></tr>
<tr><td>INFINITIVE</td><td>[pos="TO"][pos="V.*"]</td></tr>
<tr><td>HYPHENATED_WORD</td><td>[word=".*-.*"]</td></tr>
<tr><td>INDEFINITE_PRONOUN</td><td>[hw="anybody|anyone|anything|everybody|everyone|everything|nobody|none|nothing|nowhere|somebody|someone|something"]</td></tr>
<tr><td>INFINITIVE</td><td>[pos="TO"][pos="V.*"]</td></tr>
<tr><td>INTERJECTION</td><td>[pos="UH"]</td></tr>
<tr><td>LINKING_ADVERB</td><td>[hw="then|so|anyway|though|however|e\.?g\.?|i\.?e\.?|therefore|thus|nevertheless|nonetheless" pos="R.*"]</td></tr>
<tr><td>LONG_WORD</td><td>[word=".{15,}"]</td></tr>
<tr><td>MODAL</td><td>[pos="MD"]</td></tr>
<tr><td>NEGATION</td><td>[word="not|.*n't|no|neither|nowhere|never|nor|none|nobody|nothing"]</td></tr>
<tr><td>NOMINALIZATION</td><td>[word=".{3,}(tion|tions|ment|ments|ness|nesses|ity|ities)"]</td></tr>
<tr><td>NOUN</td><td>[pos="N.*"]</td></tr>
<tr><td>NUMBER</td><td>[pos="M.*"]</td></tr>
<tr><td>PARTICLE</td><td>[pos="RP"]</td></tr>
<tr><td>PASSIVE</td><td>[pos="VB[^0].*"][pos="R.*"]{0,3}[pos="V.N"]</td></tr>
<tr><td>PAST_TENSE</td><td>[pos="V.D.?"]</td></tr>
<tr><td>PAST_PARTICIPLE</td><td>[pos="V.N"]</td></tr>
<tr><td>PERFECT_INFINITIVE</td><td>[pos="TO"][pos="VH.*"][pos="V.N"]</td></tr>
<tr><td>PHRASAL_VERB</td><td>[pos="VV."][pos="PP.*"]{0,1}[pos="RP"]</td></tr>
<tr><td>PLACE_ADVERB</td><td>[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"]</td></tr>
<tr><td>PREPOSITIONAL_PHRASE</td><td>[pos="I.*|CS"][pos="J.*|PP.*|CC|D.*|RR|M.*|GE|N.*"]{0,5}[pos="N.*"]</td></tr>
<tr><td>PRESENT_PARTICIPLE</td><td>[pos="V.GK?"]</td></tr>
</table>

<table>
<tr><td>PRESENT_TENSE</td><td>[pos="V.Z"]</td></tr>
<tr><td>PRONOUN</td><td>[pos="P.*"]</td></tr>
<tr><td>PROPER_NOUN</td><td>[pos="NP.*"]</td></tr>
<tr><td>REFLEXIVE_PRONOUN</td><td>[hw=".*sel(f|ves)" pos="P.X."]</td></tr>
<tr><td>SHORT_WORD</td><td>[word=".{1,3}"]</td></tr>
<tr><td>SPLIT_INFINITIVE</td><td>[pos="TO"][pos="R.*"][pos="V.*"]</td></tr>
<tr><td>SUPERLATIVE</td><td>[pos="DAT|JJT|RGT|RRT"]</td></tr>
<tr><td>SWEARWORDS</td><td>[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"]</td></tr>
<tr><td>TIME_ADVERB</td><td>[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"]</td></tr>
<tr><td>VERB</td><td>[pos="V.*"]</td></tr>
<tr><td>PEOPLE</td><td>[sem="S2|S2:1|S2:2|S3|S3:1|S3:2|S4"]</td></tr>
<tr><td>MALE</td><td>[sem="S2:2"]</td></tr>
<tr><td>FEMALE</td><td>[sem="S2:1"]</td></tr>
<tr><td>SUPERNATURAL</td><td>[sem="S9"]</td></tr>
<tr><td>EMOTION</td><td>[sem="E|E1|E2|E3|E4|E4:1|E4:2|E5|E6"]</td></tr>
<tr><td>TIME</td><td>[sem="T1|T1:1|T1:1:1|T1:1:2|T1:2|T1:3|T2|T3|T4"]</td></tr>
<tr><td>PLANET</td><td>[sem="W1|W2|W3|W4|W5|L1|L2|L3"]</td></tr>
<tr><td>COLOR</td><td>[sem="O4:3"]</td></tr>
<tr><td>COLOUR</td><td>[sem="O4:3"]</td></tr>
<tr><td>BODY</td><td>[sem="B1|B2|B3"]</td></tr>
<tr><td>FOOD</td><td>[sem="F1|F2"]</td></tr>
<tr><td>TECHNOLOGY</td><td>[sem="Y1|Y2"]</td></tr>
<tr><td>MEDIA</td><td>[sem="Q4|Q4:1|Q4:2|Q4:3|K1|K2|K3|K4"]</td></tr>
</table>

14 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. 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

Lancaster University