



#LancsBox X manual

Designed for very large corpora and
advanced XML capabilities.
Try it with the British National Corpus 2014.

Citation for #LancsBox X:

Brezina, V., Platt, W. (2022). #LancsBox X 1.1.0 [software package]

Contents

1	Downloading and running #LancsBox X	3
2	Importing data.....	6
2.1	Visual summary: importing data	6
2.2	Load your corpora	6
3	KWIC tool (key word in context)	8
3.1	KWIC: An overview	8
3.2	Multiple panels.....	9
3.3	Metadata columns	10
3.4	Filters.....	10
3.5	Summary table	11
3.6	Working with subcorpora.....	12
4	Searching in #LancsBox	14
5	CLAWS tagset (C7)	18
6	USAS tagset	22
7	Definitions of smart searches.....	25
8	Glossary	28

#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.


#LancsBox: Lancaster University corpus toolbox
Download version X: suitable for large corpora such as the British National Corpus 2014

#LancsBox X for Windows

 **Download**

 **Windows**

 **Mac**

 **Linux**

❷ Run installer

Agree to security warnings on your machine – #LancsBox is safe to run – and follow the steps in the installer. Always install #LancsBox to a folder, where the tool has 'read and write' privileges such as the User folder or Desktop; On Windows, never install #LancsBox to Program Files.

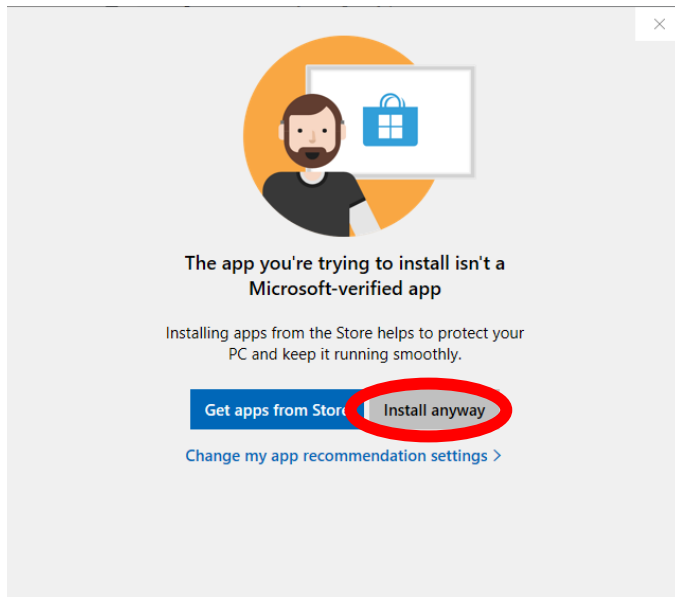
Important note: System privileges

Please follow the instruction below for your specific operating system.

Windows 10

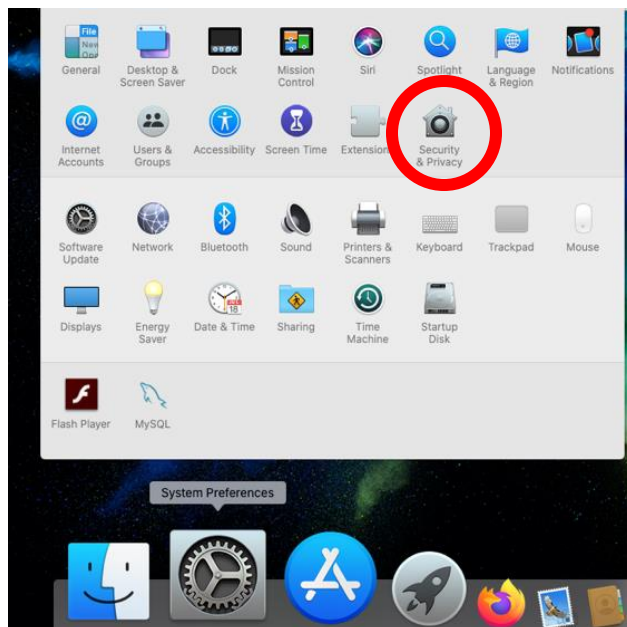
Windows 10 might display the following message.

“The app you are trying to install isn’t a Microsoft-verified app”. If this warning message appears, click on ‘Install anyway’.

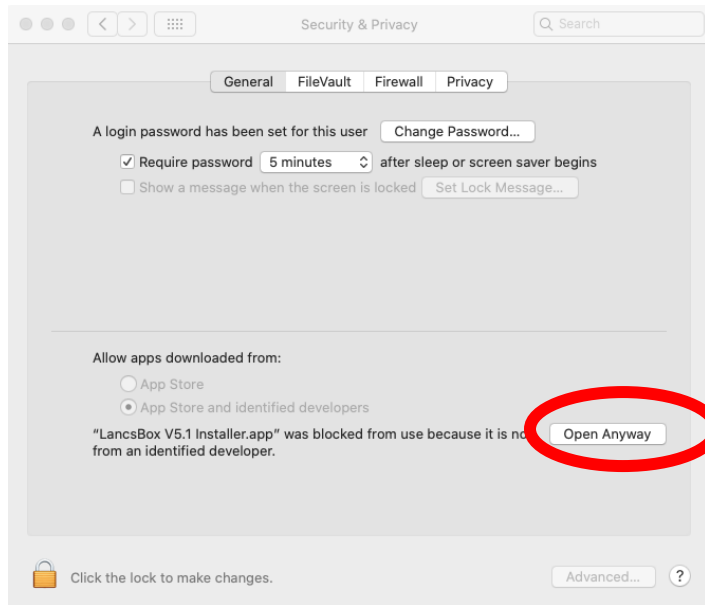


MAC

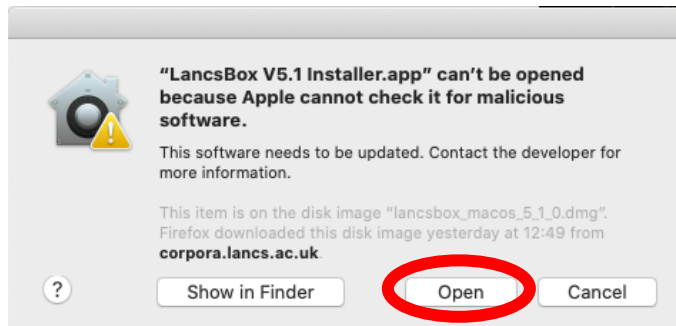
Open “System Preferences” in the dock, click on “Security & Privacy”.



Click on “Open Anyway” next to the message “LancsBox X Installer was blocked because it is not from an identified developer”.



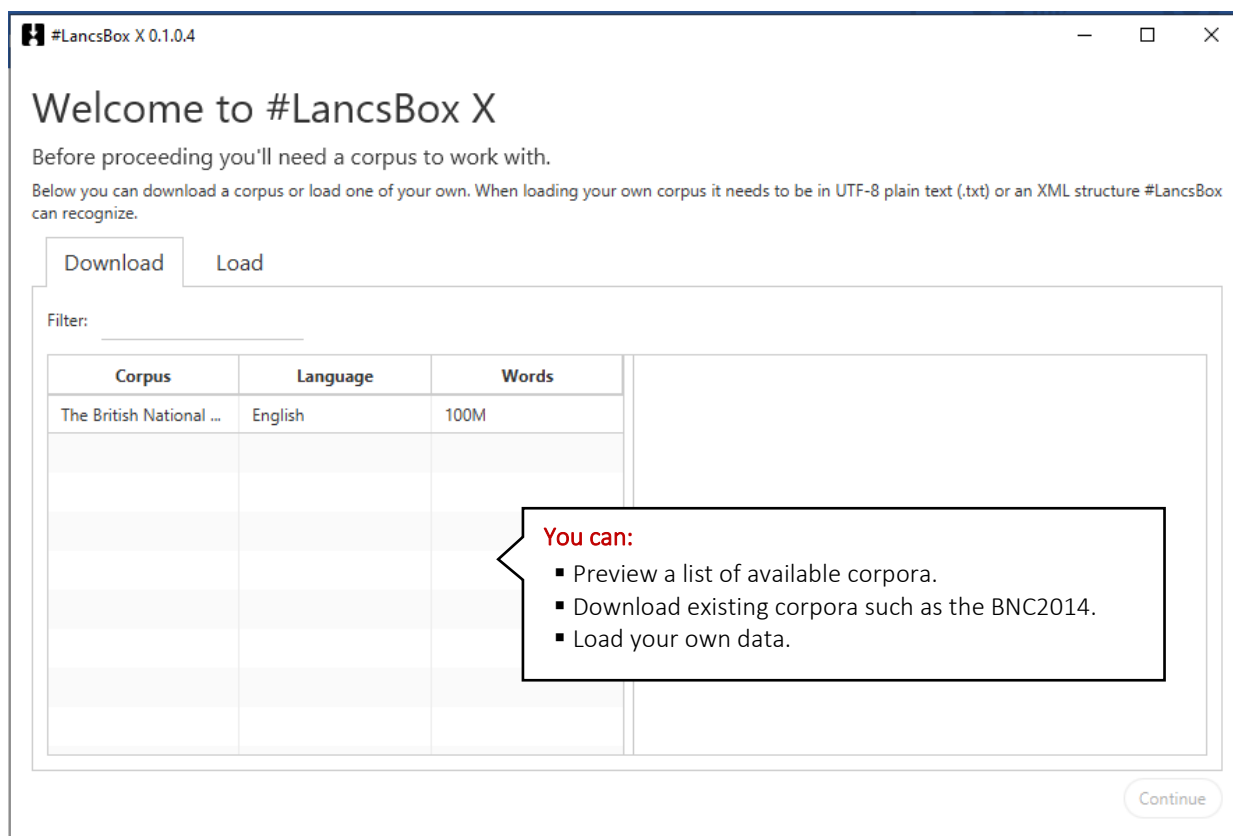
Click on “open” when the message “LancsBox X Installer.app” can’t be opened because Apple cannot check it for malicious software” is displayed in a new window.



2 Importing data

#LancsBox X is designed for very large corpora; it natively supports XML, which allows working with rich metadata. Data can be loaded and imported into #LancsBox very easily.

2.1 Visual summary: importing data



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

2.2 Load your corpora

#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><?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=</pre>

1. Prepare your data in a folder.
2. On the 'Load' tab provide information about the corpus and navigate to the data folder by clicking on 'Browse'.

Download
Load

Full name*

Short display name

Language
English

Data folder*
Browse

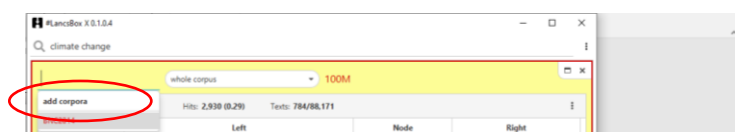
More details

Load

Continue

3. Click on 'Load'.
4. Click on 'Continue'

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



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. Single search box allows carrying 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 dropdown menu set to 'magazines', and '15M'. A 'Hits: 428 (0.29)' label is also present. A table with four columns: 'File', 'Left', 'Node', and 'Right', contains concordance lines. Annotations with callout boxes provide instructions: 'Search for a word, phrase or grammatical structure' points to the search box; 'Select corpus' points to the 'BNC2014' label; 'Select subcorpus' points to the 'magazines' dropdown; 'Left-click column header to sort. Drag to re-arrange.' points to the 'Left' column header; and 'Click the + sign to add more panels.' points to a plus icon on the right side of the table. 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

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 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.

#LancsBox X 0.1.0.4

Q PASSIVE

BNC2014whole corpus100M

PASSIVEHits: 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 Dörner	are ...	by the Board to have
O...	...ended f	e ...	in 2016, in line with
O...	appropriate and key exam	are ...	The Annual Report, taken as
O...	Code. It will continue to	be ...	during 2016. Reported to the
O...	...toring and c		the team

Progress bar

BNC2014informal speech10M

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

File	Left	Node	Right
Sp...		en't	being used
Sp...	yeah it's been they	are actually	I made a joke und
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


BNC2014academic prose20M

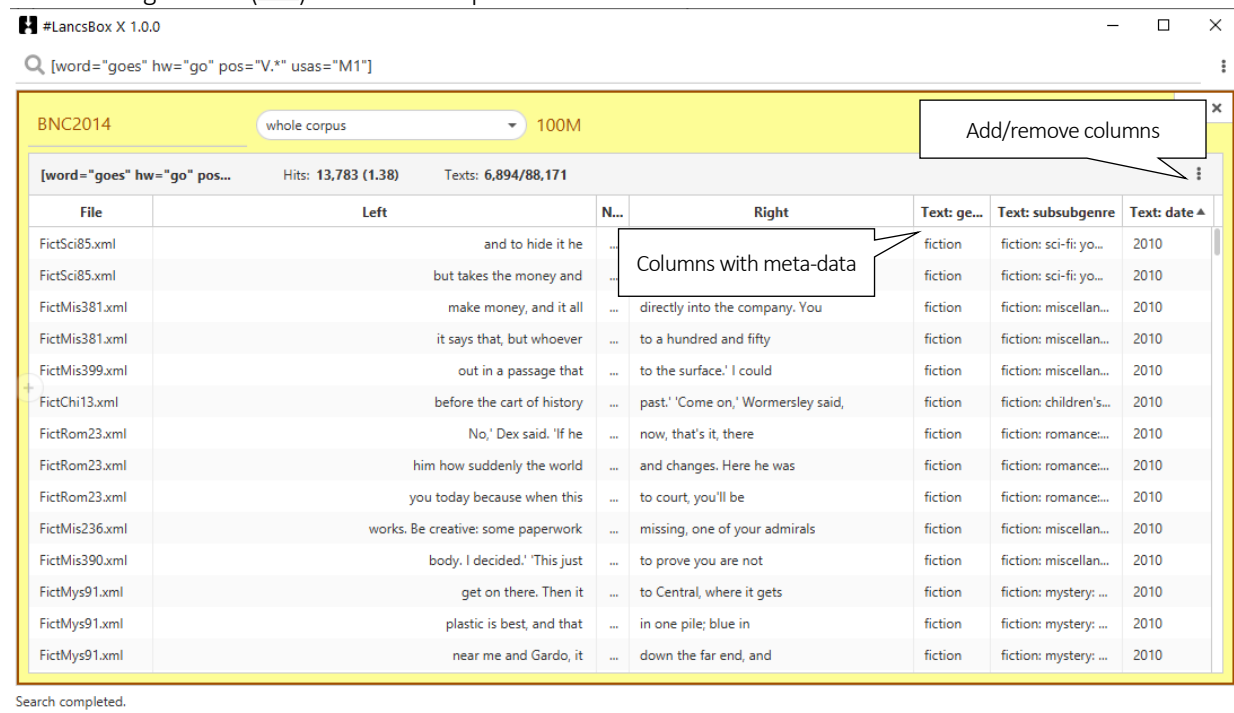
PASSIVEHits: 315,620 (161.02)Texts: 2,879/2,879

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

Searching for PASSIVE.

3.3 Metadata columns

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

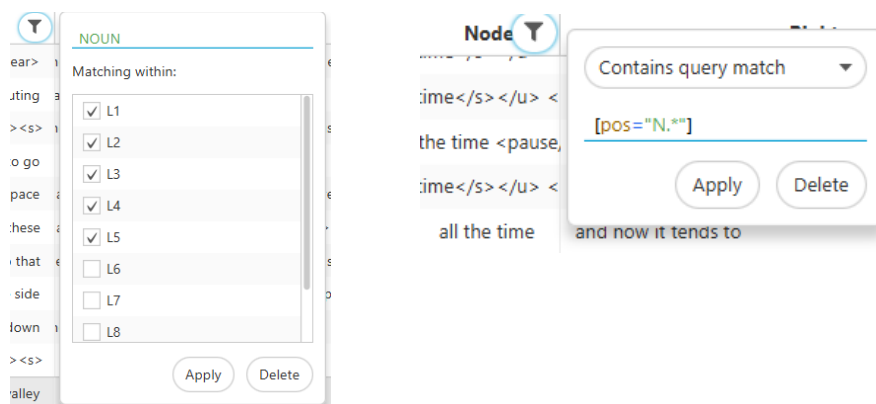
File	Left	N...	Right	Text: ge...	Text: subsubgenre	Text: date
FictSci85.xml	and to hide it he	...		fiction	fiction: sci-fi: yo...	2010
FictSci85.xml	but takes the money and	...		fiction	fiction: sci-fi: yo...	2010
FictMis381.xml	make money, and it all	...	directly into the company. You	fiction	fiction: 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

Search completed.

3.4 Filters

Powerful filters can be applied to i) linguistic and ii) metalinguistic data. Simply hover with the mouse pointer towards the right of a column header, where you wish to apply the filter.

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 be also summarised using the 'Summary table' functionality . Summary table 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  ☒ L1 ☐ L2 ☐ L3 ☐ L4 ☐ L5 ☐ L6 ☐ L7 ☐

word 

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, 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 include also a break-down by individual tags displayed in tooltips.

3.6 Working with subcorpora

#LancsBox X allows you 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” when done. Your new subcorpus will be selected.

FlencBox X1.0.1.6

all the time

Define new subcorpus Name: no restrictions

mode

- ☐ speech
- ☐ writing

genre

- ☐ academic prose
- ☐ elanguage
- ☐ fiction
- ☐ informal speech
- ☐ magazines
- ☐ newspapers
- ☐ official documents
- ☐ written-to-be-spoken

academic type

- ☐ editorial
- ☐ research article
- ☐ review article

source

- ☐ #Help: My cat's a vlogging superstar!
- ☐ #PleaseRetweet
- ☐ 11

spoken: number of speakers

- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5
- ☐ 6
- ☐ 7
- ☐ 8
- ☐ 9

spoken: inter-speaker relationships

- ☐ acquaintances
- ☐ close family, partners, very close friends
- ☐ colleagues
- ☐ friends, wider family circle
- ☐ friends and one stranger
- ☐ strangers
- ☐ NA

words

28 28,981 57,694 115,360

Number

id

- ☐ AcaHum&k1
- ☐ AcaHum&k10
- ☐ AcaHum&k11

subgenre

- ☐ academic prose: humanities
- ☐ academic prose: science

Long list of categories (searchable)

subsubgenre

- ☐ academic prose: humanities: archaeology
- ☐ academic prose: humanities: architecture
- ☐ academic prose: humanities: arts
- ☐ academic prose: humanities: arts and humanities

sample

- ☐ beginning
- ☐ composite
- ☐ end
- ☐ middle
- ☐ whole

academic publication

- ☐ book
- ☐ journal

Auto filter

date

Enabled

Start: 01/01/2010

End: 23/05/2020

Date

author

- ☐ "SWNS - CENTRE PRESS" news@swns.com
- ☐ "SWNS - NATIONAL NEWS" news@nationalnews.co.uk
- ☐ @victoriapeckham
- ☐ 1000 Women
- ☐ A.C. Davidson

OK Cancel

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 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. 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

4. 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.*" usas="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.

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"/>)

5 CLAWS tagset (C7)

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

APPG	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
VCN	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)
VVVK	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)

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

7 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 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 docking 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 obesance 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 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 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 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.*"]

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	[usas="S2 S2:1 S2:2 S3 S3:1 S3:2 S4"]
MALE	[usas="S2:2"]
FEMALE	[usas="S2:1"]
SUPERNATURAL	[usas="S9"]
EMOTION	[usas="E E1 E2 E3 E4 E4:1 E4:2 E5 E6"]
TIME	[usas="T1 T1:1 T1:1:1 T1:1:2 T1:2 T1:3 T2 T3 T4"]
PLANET	[usas="W1 W2 W3 W4 W5 L1 L2 L3"]
COLOR	[usas="O4:3"]
COLOUR	[usas="O4:3"]
BODY	[usas="B1 B2 B3"]
FOOD	[usas="F1 F2"]
TECHNOLOGY	[usas="Y1 Y2"]
MEDIA	[usas="Q4 Q4:1 Q4:2 Q4:3 K1 K2 K3 K4"]

8 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.