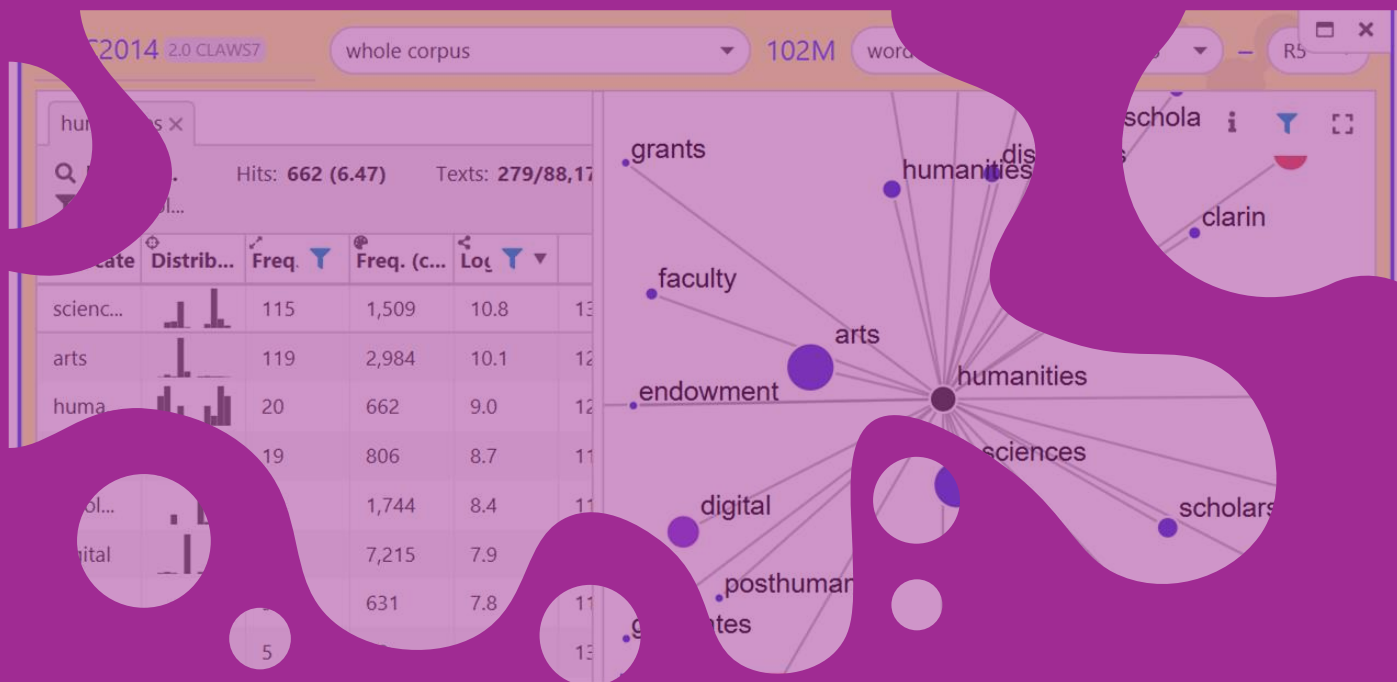


#LancsBox X



Digital Humanities

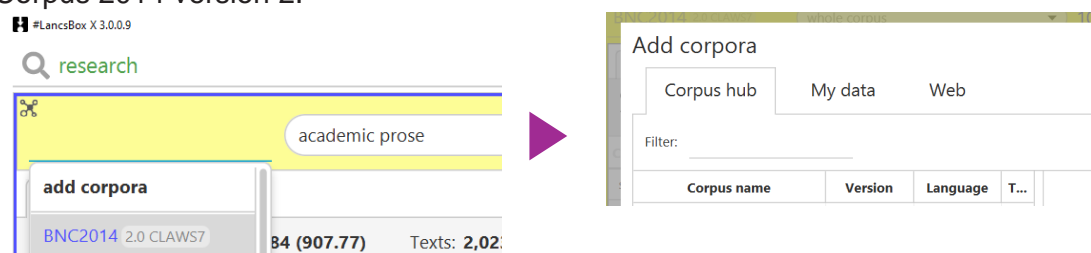
Professor Vaclav Brezina
@ Lancaster University

Starting with #LancsBox X

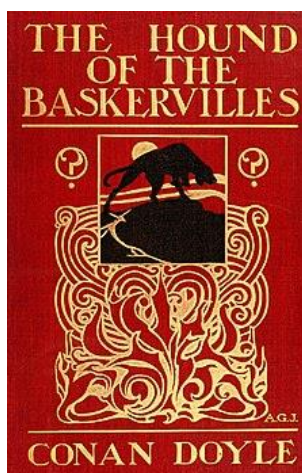
#LancsBox X is a powerful software tool for the analysis of large amounts of language. It can be used with your own data or the data provided.

The tool is very easy to use with an intuitive and flexible UI.

1. Download the most recent version of #LancsBox X from <https://lancsbox.lancs.ac.uk>
2. Go to 'add corpora' > 'Corpus Hub' and select and download the British National Corpus 2014 version 2.



Creating reality in fiction: Doyle's *The Hound of the Baskervilles*



The Hound of the Baskervilles is one of the most famous Sherlock Holmes stories by Arthur Conan Doyle (1859-1930). It appeared in *The Strand Magazine* from August 1901 to April 1902 and is set in 1889 largely on Dartmoor, Devon – Southwest of England. This is a Victorian detective story, where reason meets superstition.

“[A]void the moor in those hours of darkness when the powers of evil are exalted.

Let's explore this narrative using corpus linguistics and compare it with a large corpus!

Create your corpus

In this section, we will practice two methods of building a corpus: a manual method and an automatic method using #LancsBox X. You will then explore the data in #LancsBox X using the Text tool. The aim of this worksheet is for you to learn

- to understand the process of corpus creation
- to evaluate the quality of the data
- to understand key parameters of texts in a corpus

Text

corpus creation

Gutenberg

Web

Task 1

Build your own newspaper mini-corpus

- a) Go to Project Gutenberg and find Doyle's Hound of the Baskervilles.



Project Gutenberg
<https://www.gutenberg.org>

Project Gutenberg: Free eBooks

Project Gutenberg is a library of over 70,000 free eBooks. Choose among free epub and Kindle eBooks, download them or read them online. You will find the ...



You can also use a direct link to the text: <https://www.gutenberg.org/files/2852/2852-h/2852-h.htm>

- b) Identify the body of the text and copy (Ctrl+C) and paste (Ctrl+V) it into an empty MS Word document. Avoid copying tables of contents, copyright information and other boilerplate.

to solve it. The past and the present are within the field of my inquiry, but what a man may do in the future is a hard question to answer. Mrs. Stapleton has heard her husband discuss the problem on several occasions. There were three possible courses. He might claim the property from South America, establish his identity before the British authorities there and so obtain the fortune without ever coming to England at all, or he might adopt an elaborate disguise during the short time that he need be in London; or, again, he might furnish an accomplice with the proofs and papers, putting him in as heir, and retaining a claim upon some proportion of his income. We cannot doubt from what we know of him that he would have found some way out of the difficulty. And now, my dear Watson, we have had some weeks of severe work, and for one evening, I think, we may turn our thoughts into more pleasant channels. I have a box for *Les Huguenots*. Have you heard the De Reszkes? Might I trouble you then to be ready in half an hour, and we can stop at Marcini's for a little dinner on the way?"

THE END

*** END OF THE PROJECT GUTENBERG EBOOK THE HOUND OF THE BASKERVILLES ***

Updated editions will replace the previous one—the old editions will be renamed.

- c) Save your text document as plain MS Word document (.docx), e.g. Doyle_Hound_1902.

Congratulations! You've just created your own mini-corpus!


Task 2

Uploading your data into #LancsBox X

- a) Open #LancsBox X. From the corpus drop down menu choose 'add corpora' and select 'My data'.

Add corpora


Note: The data does not need to be in any special format – the tool accepts all major data formats (e.g. .txt, .docx., .pdf, etc.).

- b) Explore the corpus in the Text tool . Summary statistics are available when you hover with your mouse over the name of the corpus.

```
102,305,246 grammar tokens
99,949,544 space tokens
88,171 texts
Full name: The British National Corpus 2014
Version: 2.0 CLAWS7
Annotations: CLAWS7 & USAS
Folder: BNC2014_2
```

Fill in the table below with the descriptive statistics of your mini-corpus:

Corpus size – space tokens (strings separated by space)	
Corpus size – grammar tokens (identified by the tagger)	
Texts	
Lexical density (MATTR)	

- c) Tagging. When you load your own corpus into #LancsBox X, you can select 'Grammatical' and 'Semantic' tagging, which is done automatically. In the Text tool, double-click on any text and view available annotation 

Grammatical tagset: English

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

Semantic (USAS) tagset

A1	GENERAL AND ABSTRACT TERMS	B5	Clothes and personal belongings	L3	Plants
A1.1.1	General actions, making etc.	C1	Arts and crafts	M1	Moving, coming and going
A1.1.2	Damaging and destroying	E1	EMOTIONAL ACTIONS, STATES AND PROCESSES General	M2	Putting, taking, pulling, pushing, transporting &c.
A1.2	Suitability	E2	Liking	M3	Vehicles and transport on land
A1.3	Caution	E3	Calm/Violent/Angry	M4	Shipping, swimming etc.
A1.4	Chance, luck	E4	Happy/sad	M5	Aircraft and flying
A1.5	Use	E4.1	Happy/sad: Happy	M6	Location and direction
A1.5.1	Using	E4.2	Happy/sad: Contentment	M7	Places
A1.5.2	Usefulness	E5	Fear/bravery/shock	M8	Remaining/stationary
A1.6	Physical/mental	E6	Worry, concern, confident	N1	Numbers
A1.7	Constraint	F1	Food	N2	Mathematics
A1.8	Inclusion/Exclusion	F2	Drinks	N3	Measurement
A1.9	Avoiding	F3	Cigarettes and drugs	N3.1	Measurement: General
A2	Affect	F4	Farming & Horticulture	N3.2	Measurement: Size
A2.1	Affect:- Modify, change	G1	Government, Politics and elections	N3.3	Measurement: Distance
A2.2	Affect:- Cause/Connected	G1.1	Government etc.	N3.4	Measurement: Volume
A3	Being	G1.2	Politics	N3.5	Measurement: Weight
A4	Classification	G2	Crime, law and order	N3.6	Measurement: Area
A4.1	Generally kinds, groups, examples	G2.1	Crime, law and order: Law and order	N3.7	Measurement: Length & height
A4.2	Particular/general; detail	G2.2	General ethics	N3.8	Measurement: Speed
A5	Evaluation	G3	Warfare, defence and the army; weapons	N4	Linear order
A5.1	Evaluation:- Good/bad	H1	Architecture and kinds of houses and buildings	N5	Quantities
A5.2	Evaluation:- True/false	H2	Parts of buildings	N5.1	Entirety; maximum
A5.3	Evaluation:- Accuracy	H3	Areas around or near houses	N5.2	Exceeding; waste
A5.4	Evaluation:- Authenticity	H4	Residence	N6	Frequency etc.
A6	Comparing	H5	Furniture and household fittings	O1	Substances and materials generally
A6.1	Comparing:- Similar/different	I1	Money generally	O1.1	Substances and materials generally: Solid
A6.2	Comparing:- Usual/unusual	I1.1	Money: Affluence	O1.2	Substances and materials generally: Liquid
A6.3	Comparing:- Variety	I1.2	Money: Debts	O1.3	Substances and materials generally: Gas
A7	Definite (+ modals)	I1.3	Money: Price	O2	Objects generally
A8	Seem	I2	Business	O3	Electricity and electrical equipment
A9	Getting and giving; possession	I2.1	Business: Generally	O4	Physical attributes
A10	Open/closed; Hiding/Hidden; Finding; Showing	I2.2	Business: Selling	O4.1	General appearance and physical properties
A11	Importance	I3	Work and employment	O4.2	Judgement of appearance (pretty etc.)
A11.1	Importance: Important	I3.1	Work and employment: Generally	O4.3	Colour and colour patterns
A11.2	Importance: Noticeability	I3.2	Work and employmeny: Professionalism	O4.4	Shape
A12	Easy/difficult	I4	Industry	O4.5	Texture
A13	Degree	K1	Entertainment generally	O4.6	Temperature
A13.1	Degree: Non-specific	K2	Music and related activities	P1	Education in general
A13.2	Degree: Maximizers	K3	Recorded sound etc.	Q1	LINGUISTIC ACTIONS, STATES AND PROCESSES;
A13.3	Degree: Boosters	K4	Drama, the theatre and showbusiness	COMMUNICATION	
A13.4	Degree: Approximators	K5	Sports and games generally	Q1.1	LINGUISTIC ACTIONS, STATES AND PROCESSES;
A13.5	Degree: Compromisers	K5.1	Sports	COMMUNICATION	
A13.6	Degree: Diminishers	K5.2	Games	Q1.2	Paper documents and writing
A13.7	Degree: Minimizers	K6	Childrens games and toys	Q1.3	Telecommunications
A14	Exclusivizers/ particularizers	L1	Life and living things	Q2	Speech acts
A15	Safety/Danger	L2	Living creatures generally	Q2.1	Speech etc:- Communicative
B1	Anatomy and physiology			Q2.2	Speech acts
B2	Health and disease				
B3	medicines and medical treatment				
B4	Cleaning and personal care				

Q3	Language, speech and grammar	S7.3	Competition	X3.4	Sensory:- Sight
Q4	The Media	S7.4	Permission	X3.5	Sensory:- Smell
Q4.1	The Media:- Books	S8	Helping/hindering	X4	Mental object
Q4.2	The Media:- Newspapers etc.	S9	Religion and the supernatural	X4.1	Mental object:- Conceptual object
Q4.3	The Media:- TV, Radio and Cinema	T1	Time	X4.2	Mental object:- Means, method
S1	SOCIAL ACTIONS, STATES AND PROCESSES	T1.1	Time: General	X5	Attention
S1.1	SOCIAL ACTIONS, STATES AND PROCESSES	T1.1.1	Time: General: Past	X5.1	Attention
S1.1.1	SOCIAL ACTIONS, STATES AND PROCESSES	T1.1.2	Time: General: Present; simultaneous	X5.2	Interest/boredom/excite
S1.1.2	Reciprocity	T1.1.3	Time: General: Future		d/energetic
S1.1.3	Participation	T1.2	Time: Momentary	X6	Deciding
S1.1.4	Deserve etc.	T1.3	Time: Period	X7	Wanting; planning; choosing
S1.2	Personality traits	T2	Time: Beginning and ending	X8	Trying
S1.2.1	Approachability and Friendliness	T3	Time: Old, new and young; age	X9	Ability
S1.2.2	Avarice	T4	Time: Early/late	X9.1	Ability:- Ability, intelligence
S1.2.3	Egoism	W1	The universe	X9.2	Ability:- Success and failure
S1.2.4	Politeness	W2	Light	Y1	Science and technology in general
S1.2.5	Toughness; strong/weak	W3	Geographical terms	Y2	Information technology and computing
S1.2.6	Sensible	W4	Weather	Z0	Unmatched proper noun
S2	People	W5	Green issues	Z1	Personal names
S2.1	People:- Female	X1	PSYCHOLOGICAL ACTIONS, STATES AND PROCESSES	Z2	Geographical names
S2.2	People:- Male	X2	Mental actions and processes	Z3	Other proper names
S3	Relationship	X2.1	Thought, belief	Z4	Discourse Bin
S3.1	Relationship: General	X2.2	Knowledge	Z5	Grammatical bin
S3.2	Relationship: Intimate/sexual	X2.3	Learn	Z6	Negative
S4	Kin	X2.4	Investigate, examine, test, search	Z7	If
S5	Groups and affiliation	X2.5	Understand	Z8	Pronouns etc.
S6	Obligation and necessity	X2.6	Expect	Z9	Trash can
S7	Power relationship	X3	Sensory	Z99	Unmatched
S7.1	Power, organizing	X3.1	Sensory:- Taste		
S7.2	Respect	X3.2	Sensory:- Sound		
		X3.3	Sensory:- Touch		

Task 3

Creating a corpus automatically in #LancsBox X

#LancsBox X offers a feature, which allows automatic download of texts from the web. In this task, you will explore this feature. Imagine that you want to download a corpus made of Wikipedia articles related to the Hound of the Baskervilles.

- a) Locate your initial (seed) website that includes the landing page about linguistics: https://en.wikipedia.org/wiki/The_Hound_of_the_Baskervilles From the corpus drop down menu choose 'add corpora' and select 'Web'.

Add corpora

Corpus hub My data Web

Short display name

Language English

Initial URL(s)* https://en.wikipedia.org/wiki/The_Hound_of_the_Baskervilles

Limits

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

Tagging Grammatical Semantic

▼ Tagging options

Create corpus Close

1. Name your corpus

2. Paste URL(s)

3. Click on "Create corpus"

- b) Choose a different corpus name e.g. 'Wiki_Hound_tagged'. Repeat the process with Grammatical and Semantic tagging on

Tagging Grammatical Semantic

Select the appropriate tagging. If you are using this feature for the first time, the appropriate taggers and language models will need to be downloaded.

Task 4

Exploring your Wikipedia corpus in #LancsBox X

- a) Explore the size of your corpus (see instructions in Task 2b). Fill in the table below with the descriptive statistics of your Wikipedia corpus:

Corpus size – space tokens (strings separated by space)	
Corpus size – grammar tokens (identified by the tagger)	
Texts	

- b) How many times do different words related to the book occur in the corpus?

Search term	Absolute (relative) frequency	No. of texts
Holmes		
hound		
Doyle		
...		



Collocations in context with GraphColl

In this worksheet, you will explore collocation graphs and networks using the GraphColl tool in #LancsBox X. The GraphColl tool identifies collocations and displays them in a table and as a collocation graph or network. It can be used 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.

GraphColl

collocation graphs

collocation networks

Task 5

Finding collocates

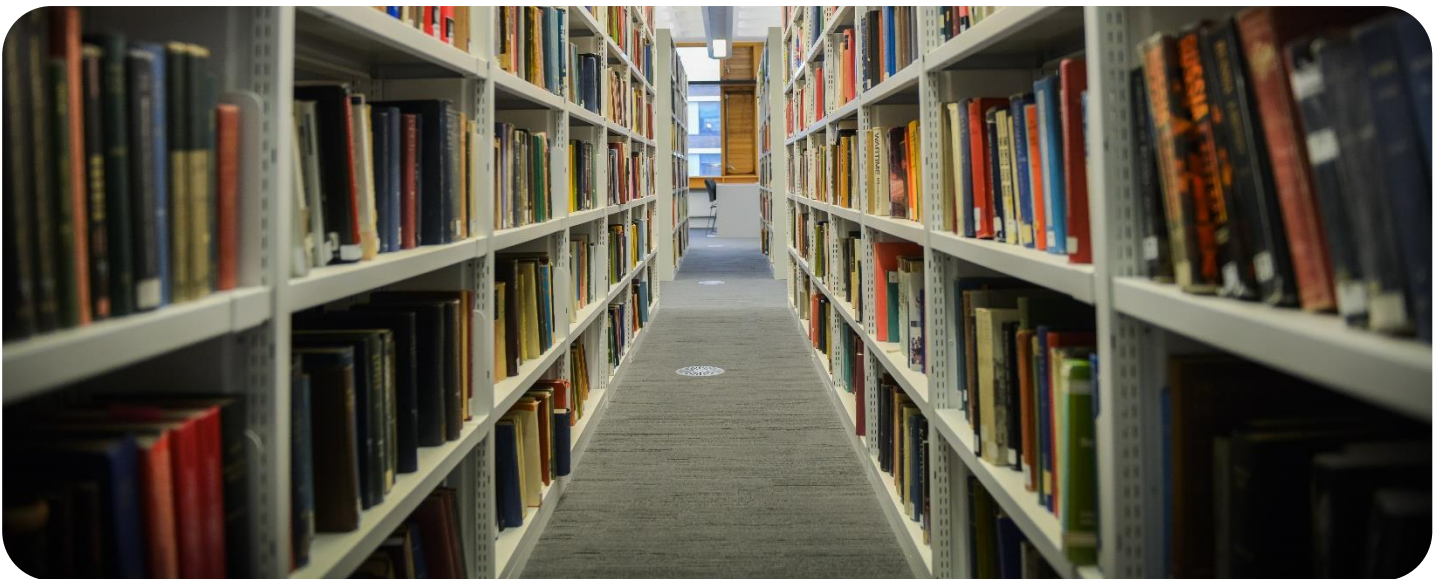
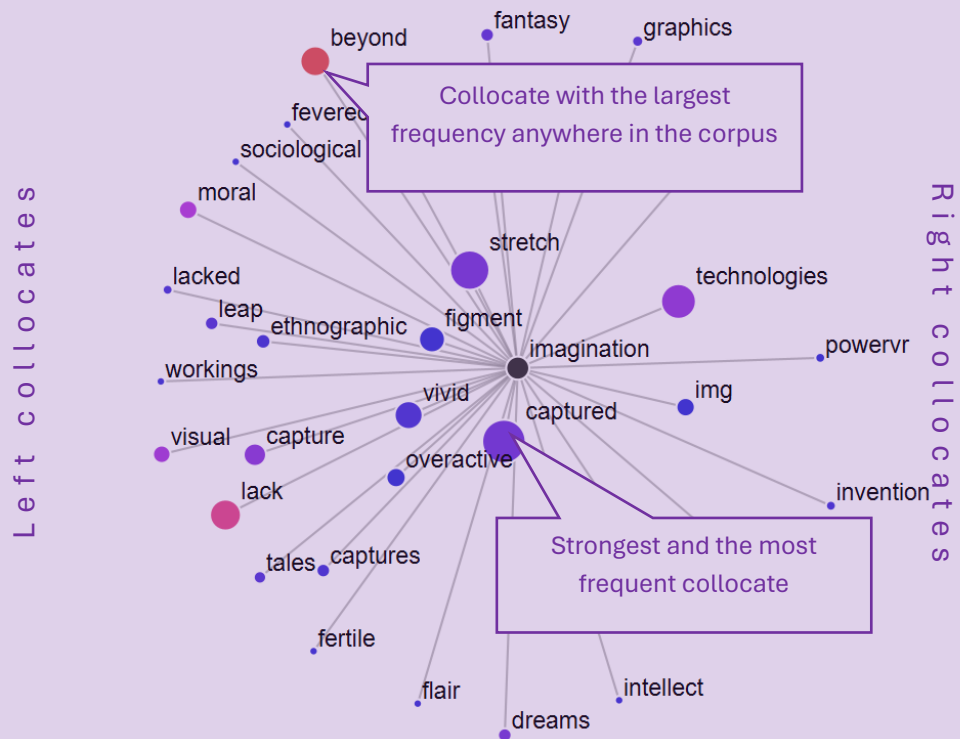
In this task, you will practice finding collocates and interpreting collocation statistics. Go to the GraphColl tool in #LancsBox X, select your Hound of the Baskervilles corpus, and search for the expressions in the table below.

Note down top collocates according to logDice and the collocation frequency.

Search term	Top 3 log Dice collocates	Most frequent collocate
moor		
hound		
VERB		

Collocation graph

A collocation graph shows the relationship between a node, which is in the middle of the graph, and its collocates, which are displayed around the node. The closer the collocate is to the node, the stronger the association. The position of the collocates indicates the position in the text, before or after the node, while the size of the collocate reflects the frequency of co-occurrence. Finally, the colour indicates the frequency of the word anywhere in the corpus on the scale from blue (small) to red (large).



Keywords and concordances: Words and KWIC

In this section, you will explore words important in a particular corpus when compared to a reference corpus. These are called *keywords*. You will also explore contexts in which these words occur using the KWIC tool, which produces concordances. You will learn how to:

- Create a keyword list.
- Understand keyword statistics.
- Produce and interpret concordances.

Wordlists

keywords

concordances

Task 7

Understanding keywords

Keywords are words that occur with a considerably higher frequency in a given (sub)corpus compared to a reference (sub)corpus. In this task, you will explore keywords in the *Hound of the Baskervilles* corpus compared to the Fiction subcorpus of the BNC2014.

1. First create a wordlist based on the *Hound of the Baskervilles* corpus and note down the first three

words: _____

2. Click on the keyword icon  and select **BNC2014 fiction** as a reference corpus.


3. Note down the top 10 keywords, i.e. words typical of the *Hound of the Baskervilles* corpus _____

4. Write a short summary of the story using all ten keywords. Underline these in your text. For further assistance see https://en.wikipedia.org/wiki/The_Hound_of_the_Baskervilles

Task 8

Understanding concordances

In this task, you will practice different search options in the KWIC tool. Go to the KWIC tool in #LancsBox X, select the *Hound of the Baskervilles* corpus, and search for the expressions in the table below. Note down their frequencies..

Tip: Hover your mouse over the results of a search (e.g. number of occurrences) and click on the “Copy to Clipboard” symbol: . Then, paste the data in the table below using the keyboard shortcut **CTRL + V** on Windows, **Command ⌘ + V** on macOS.

Hits: 571 (0.29)  Texts: 267/50,210

Copy

Search term	Occurrences (per 1M)
dark	
dark*	
dark and *	
*ment	
ADVERB dark	
TIME	
[hw="walk"]	
[sem="G2.2-. *"]	

Read the concordance lines, sort them according to the left (L) and the right (R) context. What insights do these bring into the story?



Developed at
**Lancaster
University**



#LancsBox X in your research

In this section, you will design a small study in the area of your interest. The focus is on:

- Conceptual grounding.
- Research question(s).
- Operationalization and study design.
- Data collection.
- Data analysis.

Research

design

methodology

Task 9

Designing a corpus study

In this task, you will design a mini-study. First, think about a topic in the area of your interest.

Topic: _____

Then, think of a specific question you would like to ask:

Specific question: _____

Is it a yes/know question? If no, think of an aspect of your question that can be formulated as a yes/no question.

Yes/No RQ: _____

Data: Is there an available corpus that can be used to answer the RQ? Yes/No _____

If no, can the data be obtained online? URL _____

Operationalization:

#LancsBox X tools: _____

Search terms: _____

Comparisons: _____

Possible challenges: _____