



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MT
MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Corpus Linguistics and Statistics: Practical applications

Dr. Vaclav Brezina



2018

Author:

Strategic Project of TBU in Zlín, reg. no. CZ.02.2.69/0.0/0.0/16_015/0002204

Mgr. Václav Březina, Ph.D.

Lancaster University, UK

v.brezina@lancaster.ac.uk



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education



CONTENTS

CONTENTS	3
INTRODUCTION	4
1 CORPUS LINGUISTICS: BEYOND THE BASICS AND PRACTICAL APPLICATIONS	5
1.1 CORPORA	
1.2 IMAGE: TYPES OF LINGUISTIC DATA	
1.3 TYPES OF CORPORA	
1.4 CORPUS TECHNIQUES	
1.5 IMAGE: TYPES OF CORPUS TECHNIQUES	
2 CORPUS TECHNIQUES I: EXPLORING CORPORA WITH #LANCSBOX	8
2.1 STARTING WITH #LANCSBOX	
2.2 IMAGE: DOWNLOADING AND IMPORTING CORPORA	
2.3 SEARCHING IN #LANCSBOX: KWIC AND WHELK	
2.4 ADVANCED SEARCHES	
2.5 IMAGE: SEARCHES AT DIFFERENT LEVELS	
2.6 USING CONCORDANCE LINES TO DESCRIBE A MEANING OF A WORD	
2.7 IMAGE: SORTING CONCORDANCE LINES	
2.8 FILTERS (ADVANCED USERS)	
2.9 REGULAR EXPRESSIONS (ADVANCED USERS)	
2.10 OVERVIEW OF REGULAR EXPRESSIONS.	
3 CORPUS TECHNIQUES II: BUILD YOUR OWN CORPUS.....	14
4 CORPUS TECHNIQUES III: FINDING COLLOCATIONS WITH #LANCSBOX	18
5 L1 AND L2 CORPORA AND LANGUAGE LEARNING.....	22
5.1 THE NEW GENERAL SERVICE LIST	
5.2 IMAGE: DAILY MAIL ARTICLE ON THE NEW-GSL	
5.3 LANCSEX AND THE NEW GENERAL SERVICE LIST	
6 CQPWEB AND THE BNC.....	31
6.1 IMAGE: BNC 1994 AND BNC 2014	
GLOSSARY OF CORPUS TERMINOLOGY	35
REFERENCES AND RECOMMENDED READINGS.....	37





INTRODUCTION

This study material accompanies a two-day workshop on Corpus linguistics and Statistics, arranged by Tomáš Baťa University in Zlín. It covers a series of topics in corpus linguistics and the application of corpus techniques in studies of discourse and language learning and teaching.

The material draw on the author's own research experience with corpus design and analysis; examples used come from the research carried out at The ESRC Centre of Corpus Approaches to Social Science, Lancaster University. The ESRC Centre for Corpus Approaches to Social Science, Lancaster University, UK is one of the world's leading research centres in corpus linguistics and was awarded the Queen's Anniversary Prize for Higher and Further Education in 2015. The centre specializes in applying corpus linguistic techniques to a number of research areas including sociolinguistics, education and social sciences.

Specific references to research publications are provided throughout this material. Recommended further readings are listed at the end.

A practical note: For easier orientation in the text, the 'Task' symbol () appears in the margin where a practical Task is introduced.



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education





1 CORPUS LINGUISTICS: BEYOND THE BASICS AND PRACTICAL APPLICATIONS

Let's start with a simple definition of corpus linguistics.

Corpus linguistics is a versatile methodology of computational analysis of language, which can be used for the study of grammar (e.g., Biber et al., 1999), discourse (e.g., Baker, 2006), different social actors and processes (Baker et al., 2008) as well as in language learning and teaching (e.g., Aijmer, 2009). Corpus linguistics is a scientific approach to language, which follows the tradition of empirical investigation, that is, investigation in which the focus is on the collection and interpretation of data (McEnery and Hardie, 2011). Somewhat symbolically, the importance of data for corpus linguistics is evident from the fact that corpus linguistics has a term that refers to data, namely corpus, in its name.

Brezina & Gablasova (2018: 595)

The focus in this quote is on the methodological aspect of the discipline, which uses modern computational technology to analyse large amounts of naturally-occurring language, which is compiled in a form of corpora (*sg.* corpus). Corpora can be analysed for linguistic and social purposes, an area of research that has been pioneered by the The ESRC Centre of Corpus Approaches to Social Science at Lancaster University. Given this basic understanding of corpus linguistics as primarily a methodology (rather than theory of language; cf. McEnery & Hardie 2011), the major focus of this material will be on practical tool descriptions and exercises. These will allow the reader (workshop participant) to apply the techniques learnt to their own data in their own research context.

1.1 Corpora

Corpora as a source of linguistic and social evidence represent natural type of data, both spoken and written, that appear in a variety of contexts from informal conversation to essays,



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education





novels, academic books and social media. Figure 1 below adapted from Ellis, 2015 shows different types of linguistic evidence; as mentioned above, corpora represent data on language use that is usually not elicited but contains collected examples of natural use of language ‘in the wild’.

1.2 Image: Types of linguistic data

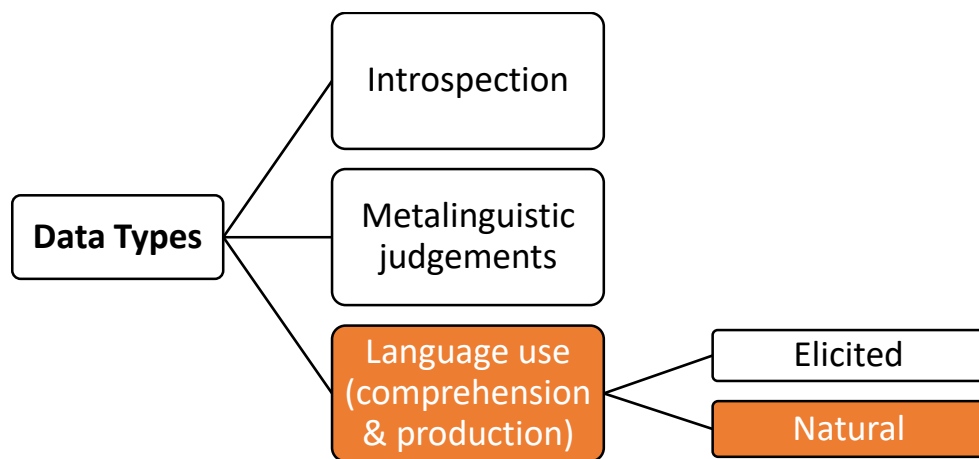


Figure 1. Type of linguistic data

1.3 Types of corpora

Corpora can be classified in different ways. There are three major criteria according to which corpora can be classified.

- According to the mode of communication: spoken, written and multimodal
- According to scope and variety: general and specialised
- Type of language represented: native-speaker (L1) and non-native speaker (L2, learner)

The user always needs to select a corpus which is suitable for answering their research question(s); if no such corpus is available the user can build their own corpus (see section 3).

1.4 Corpus techniques

Corpus linguistics uses a range of quantitative (statistical) and qualitative techniques for the analysis of data compiled in corpora. Four ‘core’ methods focus on i) frequency of words,





phrases and grammatical structures, ii) their distribution, iii) use and iv) co-occurrence. Frequency is typically displayed as part of concordances or in frequency lists. Information about words' distribution is provided through various techniques investigating dispersion. A word's use in context is typically explored using concordance lines, while its co-occurrence with other words is analysed through collocation.

1.5 Image: Types of corpus techniques

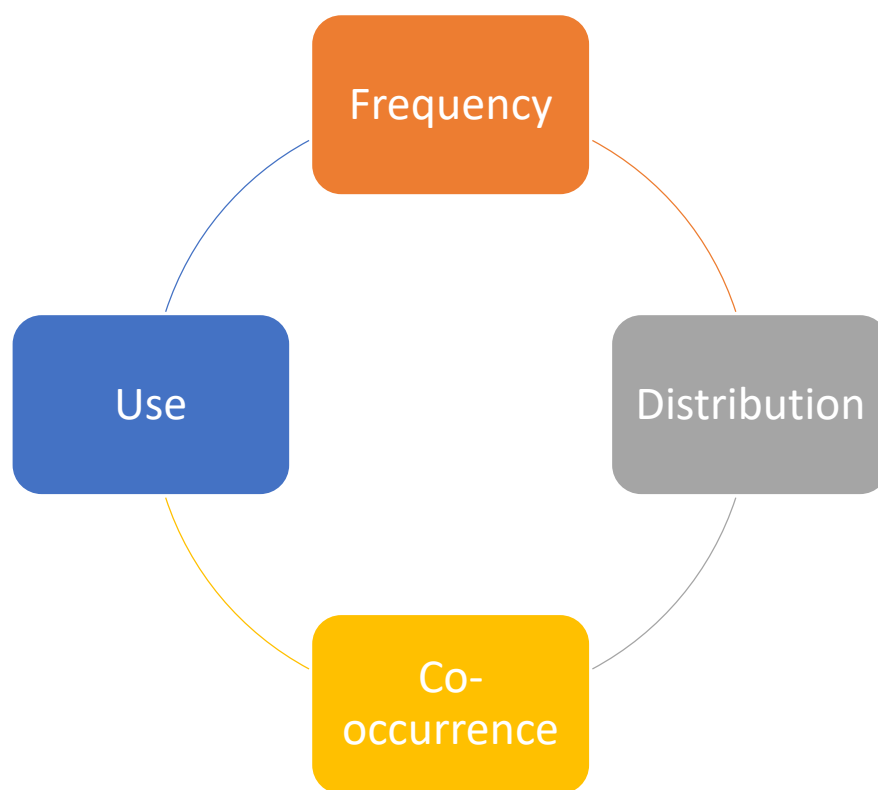


Figure 2. 'Core' corpus techniques





2 CORPUS TECHNIQUES I: EXPLORING CORPORA WITH #LANCSBOX

In this series of practical tasks #LancsBox, a new software package for the analysis and visualisation of corpus data will be introduced. #LancsBox was developed at Lancaster University and is distributed for free for non-commercial (research and teaching) purposes.

#LancsBox is available for download from: <http://corpora.lancs.ac.uk/lancsbox>

How to cite #LancsBox?

Brezina, V., McEnery, T., & Wattam, S. (2015). *Collocations in context: A new perspective on collocation networks*. *International Journal of Corpus Linguistics*, 20(2), 139-173.

2.1 Starting with #LancsBox

The starting point for using the tool is the 'Corpora' tab where data is loaded and further processed (POS-tagged etc.). There are two options for loading corpora: i) loading local corpora from your computer and ii) downloading corpora provided by #LancsBox.

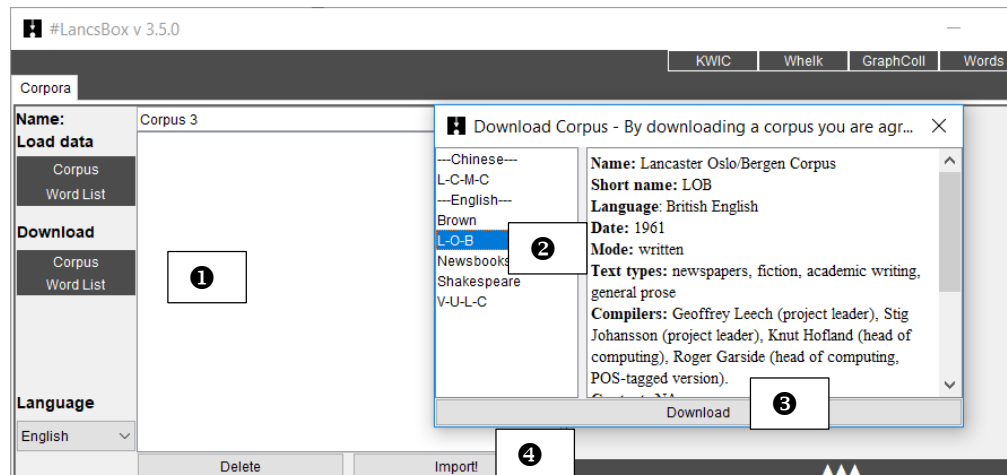


Task 1. On the Corpora tab, click on 'Corpus' under 'Download'. In a pop-up window select L-O-B, click 'Download' and when the corpus is downloaded, click 'Import'. Wait for the corpus to import – this might take some time... (see the instructional image below).





2.2 Image: Downloading and importing corpora



2.3 Searching in #LancsBox: KWIC and Whelk

The KWIC tool generates a list of all instances of a search term in a corpus in the form of a concordance. The Whelk tool provides information about how the search term is distributed across the corpus files.

These tools can be used to:

- Find the frequency of a word or phrase in a corpus.
- Find, sort and filter the examples of the use of a word or phrase in the corpus.
- Find complex linguistic structures such as the passives, split infinitives etc. using ‘smart searches’.



Task 2. Search for the following expressions in LOB and note down their frequencies and distributions in texts:

Search term	Occurrences (per 10k)	Texts
research		
research*		
research and *		
NOUNS		





PASSIVES		
>1950&<1960		

Note an asterisk (*) in the searches is used as a special character (wildcard) meaning zero or more characters.

2.4 Advanced searches

#LancsBox allows searching at different levels: i) plain text, ii) headword, iii) POS (part of speech) and iv) user-defined tags. These advanced searches are activated by clicking on the down arrow inside the search box.

2.5 Image: Searches at different levels

go	Headword
V*	POS

Text level empty → no constraint.

Headword is go.

AND

POS is any verbal use.



Task 3. Search for the following expressions in LOB and note down their frequencies and distributions in texts; use the advanced search features in #LancsBox.

Search term	Notes	Occurrences (per 10k)	Texts
go	as headword		
play	as noun (N*)		
light	as adjective (J*)		

2.6 Using concordance lines to describe a meaning of a word

Concordance lines can be used to describe the meaning of a word. Concordance lines can be sorted according to the immediately preceding (left) or immediately following (right) context. In #LancsBox, simply click on header of the concordance which says 'Left' or 'Right'.



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education





2.7 Image: Sorting concordance lines

Index	File	Left	Node	Right
899	J_Acad_writir	to Manchester he and Geiger, using Geiger's	new	15a-particle counting techniques, were able to m
449	E_Skills.bt	beaches, and the Pindos(2nd class). A	new	1st-class hotel, the Hibiscus, with private beach,
496	E_Skills.bt	the milk is filled in to the	new	250 ml. jars, with a 55 mm.
468	E_Skills.bt	12in x 4in section, set out on	new	2ft thick slabs, the channels being bolted
324	C_Press_revi	but it is now distinguished by a	new	Abigail(Elizabeth Vaughan) who tackled her very
150	A_Press_rep	AFTER making a hesitant start to the	New	Account, stock markets have taken some encour
580	F_Pop_lore.b	open up into a royal road to	new	achievements and exciting discoveries. So the m
527	F_Pop_lore.b	to hurt a man's feelings. With a	new	acquaintance, it is probably best to let
1,045	K_Fiction_gei	friendly laughter in Andre2's flat and the	new	acquaintances she had made through the Frencl
978	J_Acad_writir	trust instrument or by statute. Before this	new	Act the investments authorised by statute did
979	J_Acad_writir	twenty years before the passing of the	new	Act, the second object had not been
42	A_Press_rep	to appeal against their detention under the	new	Act. An officer of the council said
553	F_Pop_lore.b	he was should take medical advice immediately.	New	activities, of whatever kind, are free from
465	E_Skills.bt	surreptitiously, I was shown the bathroom— a	new	addition, painted blue and adorned with a
762	G_Belle_Jett	with Cuba, but the advent of a	new	administration certainly provides a new opportun
771	H_Misc_non	therefore had to consider:(1) whether any	new	advice could be brought effectively to the
1,046	K_Fiction_gei	of state in the hands of his	new	adviser. Queen Augusta, who had hitherto seen
1,036	K_Fiction_gei	a few of his associates in the	New	Africa Group became involved in this stream
746	G_Belle_Jett	Overseas Service Patterns of Government in the	New	Africa Is a Party System Possible in
1,037	K_Fiction_gei	Vaughan, Hadley, Prebble, Lambert-Price— the	New	Africa lot—"" Do they confide in you?
72	A_Press_rep	symbol. In at least one of the	new	African states, Somalia, the two Communist pow
996	J_Acad_writir	use the two key advances of the	new	age extensively. The coke-fired blast furnace and



Task 4. Imagine you are a lexicographer who writes a dictionary entry. Choose ONE of the words below and describe their meaning. If the word has multiple meanings, distinguish between senses. Provide examples of use and comment on any lexico-grammatical patterns.

Words to choose from: 1) cause 2) bring about

My corpus-based dictionary

headword:

word class:

definition:

patterns:

example(s):



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education



MINISTRY OF EDUCATION,
YOUTH AND SPORTS



Task 5. Check your dictionary entry with your neighbour. Did you come up with a similar solution?

Now check the definition in [Oxford Advanced Learner's Dictionary online](#). How does your solution differ from the dictionary solution?

2.8 Filters (advanced users)

Filters are activated by right-clicking inside concordance lines. Filters are typically applied to the left or the right context. Advanced filter is called when the top bar of the concordance table is right-clicked.

2.9 Regular expressions (advanced users)

Regular expressions represent one of the most advanced aspects of searches in corpus linguistics. They represent a universal computational 'meta-language', which allows searching for any string (characters and spaces) or a combinations of strings. Beginners should feel free to skip this section.

2.10 Overview of regular expressions.

Regex	Explanation	Regex	Explanation
Word	A string of characters (case sensitive)	a{3}	Exactly 3 of a
/word/i	A string of characters (case insensitive)	a{3,}	3 or more of a
/word\p	Punctuation search: A string of characters followed by full stop (case sensitive)	a{3,6}	Between 3 and 6 of a
[abc]	A single character either a, b or c.	\d	Any digit
[^abc]	Any single character except: a, b, or c	\D	Any non-digit





[a-z]	Any single character in the range a-z	\w	Any word character (letter, number, underscore)
[a-zA-Z]	Any single character in the range a-z or A-Z	\W	Any non-word character
[0-9]	A single number in the range 0-9		
.	Any single character		
(a b)	a or b		
a?	Zero or one of a		
a*	Zero or more of a		
a+	One or more of a		

Table 1: Regular expressions





3 CORPUS TECHNIQUES II: BUILD YOUR OWN CORPUS

In this session, you will first create your mini-corpus of newspaper articles. To analyse the corpus, we will be using **#LancsBox**, a free multi-platform tool for the analysis of language developed at Lancaster University.



Task 6. Building your own newspaper mini-corpus. Follow the steps below.

1. Decide on the newspaper you want to use. Choose ONE of these:

Newspaper	Website
The Guardian	www.theguardian.com
The Telegraph	www.telegraph.co.uk
The Daily Mail	www.dailymail.co.uk
Prague Daily Monitor	praguemonitor.com

2. Choose a word or phrase that characterizes a topic of your interest (e.g. “English teaching”, “ecology”, “school tests”, “poverty”, “climate change”).

My word/phrase is: _____

3. Go to *Google* to search for articles in the selected newspaper. Include your word/phrase(s) in the search,

e.g. "climate change" site:www.theguardian.com

NB: There is no space between *site* and the web address.

4. Open the articles returned by Google (or other search engine) one-by-one and copy-paste them into a Word document. Do this with 5-10 articles. Save the file.

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



Task 7. Uploading your corpus into **#LancsBox**. Upload your own corpus into **#LancsBox**. Follow these steps:

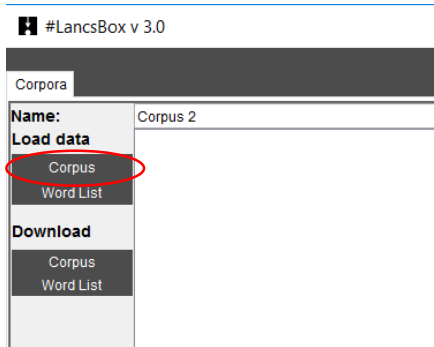


EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education





1. In the ‘Corpora’ tab, left-click on ‘Corpus’ under ‘Load data’ (see below).
2. This will allow you to browse your files and folders to select the location of your corpus. Select the file with your corpus. Enter the name of your corpus in the ‘Name’ field. Left-click ‘Import!’



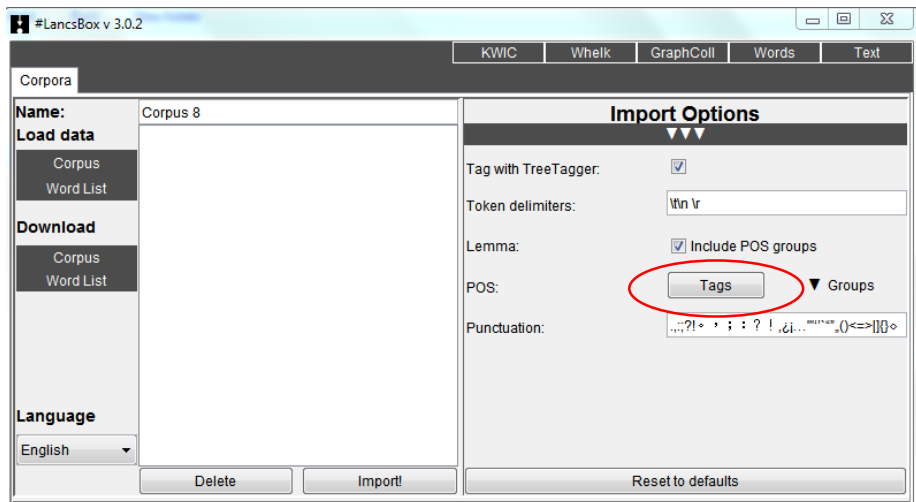
T Task 8. Using #LancsBox to explore your corpus.

1. Explore the size of your corpus and note it down:

Corpus size – tokens (running words):	
Corpus size – types (different words):	
Corpus size – lemmas (headword + POS category):	

2. Annotation – POS-tagging. When you load your own corpus into #LancsBox, it will be automatically tagged for part-of-speech categories using TreeTagger.
 - a. Check the POS categories that TreeTagger uses. In the ‘Import Options’, click on ‘Tags’ (see below).

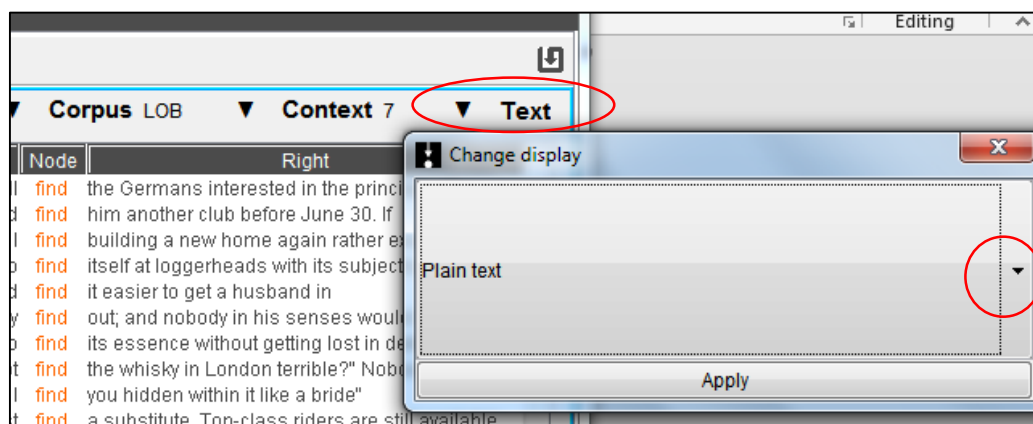




b. KWIC tool. Go to the 'KWIC' tool and search for the frequency of your word or phrase in your mini-corpus and note it down.

The frequency of my word/phrase is:

c. In the last step, your word/phrase was displayed in the KWIC view. Now go to the View option. Select the 'Text with POS' view. This will show you how your text was tagged for parts-of-speech.



Task 9. Doing concordance analysis in #LancsBox. Go to the KWIC tool in #LancsBox and search for the following expressions in your corpus and note down their frequencies and distributions in texts:

NOUNS

VERBS



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MSMT
MINISTRY OF EDUCATION,
YOUTH AND SPORTS



PASSIVES.....

- T** **Task 10.** Reflect on what type of data could you get from the Internet for constructing a corpus. What type of discourse could you find on the Internet? What type of genres? What are the possible limitations of trying to get data from the Internet? What are the possible ethical considerations?





4 CORPUS TECHNIQUES III: FINDING COLLOCATIONS WITH #LANCSBOX

The GraphColl tool identifies collocations and displays them in a table and a graph.

It can be used to:

- Find the collocates of a word, phrase or word class.
- Visualise the collocations of a word or phrase and identify connections between multiple words (collocation network); the graphs illustrate how closely the node and collocates occur together, the strength of the collocation (as measured by a particular association measure) and the position of the collocates in text in relation to the node.

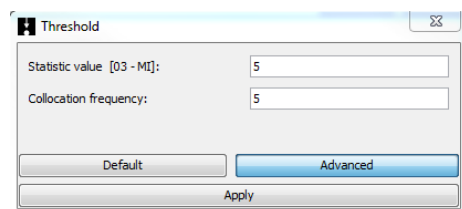


Task 11. Create graphs. Work with the LOB corpus.

- a) Build a collocation graph (first-order collocates) around the word *time* using MI score and the default settings.



- b) How many collocates does the graph display? Are all of them useful?
- c) Change the default settings as indicated in the figure below (MI = 5 and above) and search for the node *time* again.



How many results did you get this time?



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education





- d) Which of the collocates occur predominantly to the left of the node *time* and which ones to the right?

Left:

.....
.....

Right:

.....
.....

- e) Some of the collocates of *time* such as *t*, *kungo* might not be completely transparent. Use the right-click function to obtain concordances (KWIC pop-up) and explain these collocates.

t is used as

.....
.....

kungo is used as

.....
.....

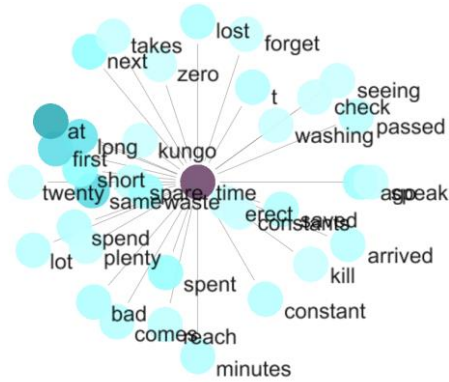
T **Task 12.** Build collocation networks and explore graphs.

- a) Go to the graph you have created in Task 11 c). It should be similar to the graph displayed in the figure below:

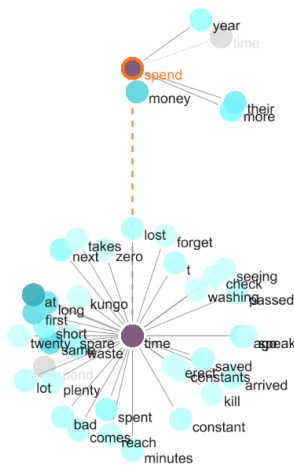


EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education





- b) Find the collocate *spend* in the graph and double click on it. You should get a collocation network similar to the one displayed below:



- c) Find the second-order collocate *money* in the graph and double click on it. Comment on the connection between *time* and *money* that you can see in the resulting graph that shows collocates around the node *money*.
- d) Repeat steps a) – d) with lemma(▼ Lemma) as the unit in the collocation graph.





T **Task 13.** Produce a range of collocation graphs and collocation networks using different corpora and experimenting with different settings (span, statistic, threshold and unit). Suggested starting points (initial nodes): *make* [all forms used as verbs], *air* and *beautiful*.



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education





5 L1 AND L2 CORPORA AND LANGUAGE LEARNING

There are many corpora that can be used for research and teaching. The following is an overview of useful corpora for the exploration of English.

Corpus	Modality	Variety	Size (words)	Availability
General				
BNC	spoken (10%) and written (90 %)	current British English	100 million	CQPweb
COCA	spoken (20%) and written (80 %)	current Ameri- can English	530 million	BYU corpora
BE06	written	current British writing	1 million	CQPweb
AM06	written	current British writing	1 million	CQPweb
English Web (En- TenTen13)	written	current interna- tional English online	20 billion	SketchEngine
Specialised				
BASE	spoken	British aca- demic speech	1.2 million	SketchEngine (free)
BAWE	written	British aca- demic writing	7 million	SketchEngine (free)
EEBO-TCP v.3	written	historical Brit- ish English (1400-1800)	1.2 billion	CQPWeb





COHA	written	historical American English (1810s-2000s)	400 million	BYU corpora
------	---------	--	-------------	-------------

Table 2: Overview of corpora

Lancaster University has embarked on a new project to create the British National Corpus 2014. The British National Corpus 2014 is a major project led by Lancaster University to create a 100 million word corpus (a large collection of ‘real life’ language) of modern-day British English. This corpus will be used by researchers to understand more about how language works and how it is evolving. Educators, dictionary compilers and the interested public will also be able to access it to find usage examples of modern published British English in different genres.



5.1 The New General Service List

Learning vocabulary is a complex process in which the learner needs to acquire both the form and a variety of meanings of a given vocabulary item. For the beginner, the main question, of course, is where to start. Arguably, the best starting point would be to look at general words. These are words that occur frequently in different texts across different genres. Brezina and Gablasova (2015) identified useful words occurring frequently in a number of different contexts in English. The result of this investigation is the New General Service List (new-GSL) that contains approximately 2,500 words. The full study has been published in the *Applied Linguistics* journal. The new-GSL is based on the analysis of frequent general words across four language corpora (LOB, BNC, BE06 and EnTenTen12) of the total size of over 12 billion running words.





The new-GSL was also featured in an article in the popular British newspaper the Daily Mail, although the paper didn't get the focus quite right (see the author's reaction to the article).

5.2 Image: Daily Mail article on the new-GSL

Forget supper and soup... it's all sex and celebrity now: How our most commonly used words have changed in 50 years

- Words such as 'sex', 'celebrity' and 'Christmas' all on the up since the 60s
- However, 'modesty', 'gentleman' and 'marriage' are less common these days

By [MARK HOWARTH](#)

PUBLISHED: 02:05, 16 September 2013 | **UPDATED:** 07:47, 16 September 2013



50

View comments

If the words we use every day reveal anything about our society, then gone are the days of handkerchiefs, handwriting, soup and suppers.

Welcome instead to the new world of mortgages and movies, celebrity and sex.

Researchers have ranked 2,500 of our most commonly used words and compared them with the same list from



Original Daily Mail article is available at: <http://www.dailymail.co.uk/news/article-2421569/Forget-supper-soup--sex-celebrity-How-commonly-used-words-changed-50-years.html>



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MSMT
MINISTRY OF EDUCATION,
YOUTH AND SPORTS



Author's reaction is available at:

Brezina, V. (2013) "It's all sex and celebrity now": Page three corpus linguistics.
<http://cass.lancs.ac.uk/?p=852>

5.3 LancsLex and the New General Service List

In this session, you will be using **LancLex, the Lancaster Vocabulary Tool** available at <http://corpora.lancs.ac.uk/vocab>. This tool analyses lexical coverage of texts, which can be used as a measure of lexical complexity. The tool differs from other lexical coverage tools (e.g. Lextutor) in that it analyses texts grammatically, distinguishing between different word classes, before comparing lexical items in text with the New General Service List (Brezina & Gablasova, 2015).

With this tool you can:

- Identify 2,500 most frequent English words in your text
- Identify specialised vocabulary
- Check lexical appropriateness of teaching materials
- Measure lexical complexity of texts for research purposes

T Task 14. Getting to know the tool: Intuition vs evidence from corpus: Read the text below and answer questions ❶ and ❷.



Russian artist Svetlana Petrova has become known for her online artwork of famous portraits featuring her big ginger cat Zarathustra.

Ahead of a new exhibition bringing the internet meme into a physical setting, the artist tells the BBC why she first created the artwork and how digital technology is helping to create new art forms.

"I lost my mother in 2008 and she left me Zarathustra. I got horrible depression after her death and for two years I was unable to do something creative. By chance a friend asked me 'why don't you make an art project with your cat because he's so funny'.

Source: Hassan, Genevieve The paintings 'made better with cats' <http://www.bbc.co.uk/news/entertainment-arts-27507950>



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education





❶ From the text, pick three nouns which you think are frequent in English. Also, pick three nouns which you think are relatively rare. Write these in the second column (“Nouns: your intuition”) of the table below.

❷ Using the online Lancaster vocabulary tool (<http://corpora.lancs.ac.uk/vocab>), check the ranking of the nouns you have picked. Enter the answers in column 3 (“new-GSL rank”).

Was your intuition correct?

Table: Nouns in the text

	Nouns: your intuition	Nouns: new-GSL rank
Frequent	1.	
	2.	
	3.	
Rare	1.	
	2.	
	3.	

LancsLex tip:

1. Go to the website and search for the nouns in the table above using the search tool.
2. The search tool provides information about the word class and the rank of the words in the new-GSL. “No matches found” means that the word does not occur among 2,500 most frequent words in English.





T Task 15. Getting to know a text: Analysing text with LancsLex.

Look at the text below. Use the LancsLex to analyse the lexical structure of the text.

- What are the most frequent words in the text? What are the infrequent words?

.....

.....

.....

.....

- What kind of words appeared on the 'Off-list' (i.e. the list of words not found in the new-GSL) and what do they tell us about the text (or corpus).

.....

.....

.....

.....

- If you are a teacher how can you use the information about word frequency to select or prepare texts for your class? (Think about whether you need to pre-teach some words or simplify the text). How could the different wordlists produced by the tool be used (e.g. in ESP classes or L2-medium content classes)?

.....

.....

.....

.....



**LancsLex tip:**

1. Go to the website and copy (Ctrl + C) the text from Materials – Texts for analyses.
2. Go to the 'Analyse' page, paste (Ctrl + V) the text in the text box and press 'Analyse text now'.
3. Explore the different analytical displays in the output.

Analyse your texts with the new-GSL

English vocabulary interactive resource

Paste the text you want to analyse into the text box below.

The first Māori settlements were mostly located around harbours or river mouths where fish and sea birds lived abundantly. New Zealand, unlike their original islands, was abundant in wild game, so the Māori used both agriculture and hunting to sustain the Iwi. One of their biggest sources of food was the moa, a large flightless bird. The moas varied in size from the height of a turkey, to 3.7 metres high. Unfortunately, this made them easy targets, and they became extinct due to over-hunting by about 1500. As a result of this, the Māori switched back to agriculture.

Gradually, the Māori dispersed themselves over New Zealand in different tribes, with different chiefs as leaders. The different tribes became more aggressive however, and inter-tribal warfare became much more frequent over time. This led to the advent of the pa (a fortified village). An average pa included ditches, banks and palisades as protection. New Zealand eventually became divided up by tribal territories which were recognised by other tribes with predominant land features (rivers, mountains, lakes). This culture remained up until the 18th Century, when the Europeans came to New Zealand.

☒ filter proper nouns ☒ filter numbers ☐ include American supplement☐ compute complexity measures**Analyse text now** **TEXT for ANALYSIS**

The first Māori settlements were mostly located around harbours or river mouths where fish and sea birds lived abundantly. New Zealand, unlike their original islands, was abundant in wild game, so the Māori used both agriculture and hunting to sustain the Iwi. One of their biggest sources of food was the moa, a large flightless bird. The moas varied in size from the height of a turkey, to 3.7 metres high. Unfortunately, this made them easy targets, and they became extinct due to over-hunting by about 1500. As a result of this, the Māori switched back to agriculture.

Gradually, the Māori dispersed themselves over New Zealand in different tribes, with different chiefs as leaders. The different tribes became more aggressive however, and inter-tribal warfare became much more frequent over time. This led to the advent of the pa (a fortified village). An average pa included ditches, banks and palisades as protection. New Zealand eventually became divided up by tribal territories which were recognised by other tribes with predominant land features (rivers, mountains, lakes). This culture remained up until the 18th Century, when the Europeans came to New Zealand.

Source: *New Zealand history* (wiki book), http://en.wikibooks.org/wiki/New_Zealand_History



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education



MINISTRY OF EDUCATION,
YOUTH AND SPORTS



- T Task 16. Modifying a text:** Imagine you are a textbook writer or a teacher who wants to create a teaching material based on real sources. Transform the text below to make it appropriate for a lower-intermediate learner group. Look at the example and follow the instruction below.

Example: Gradually, the Māori dispersed themselves over New Zealand in different tribes, with different chiefs as leaders. The different tribes became more aggressive however, and inter-tribal warfare became much more frequent over time. This led to the advent of the pa (a fortified village). An average pa included ditches, banks and palisades as protection.

The vocabulary coverage of the original text is 72%

See below an example of the modified text with lexical simplifications (highlighted).

After some time, the Māori moved to different parts of New Zealand and formed different tribes, with different chiefs as leaders. The different tribes became more aggressive however, and inter-tribal wars became much more frequent over time. This led to the introduction of the pa (a fortified village). An average pa included ditches, banks and palisades as protection.

The vocabulary coverage of the modified txt is 80%

- 1 Now, it is your turn.** Use LancsLex to help you identify less frequent words in the text below. Also note the vocabulary coverage of the original text.

The first Māori settlements were mostly located around harbours or river mouths where fish and sea birds lived abundantly. New Zealand, unlike their original islands, was abundant in wild game, so the Māori used both agriculture and hunting to sustain the Iwi. One of their biggest sources of food was the moa, a large flightless bird. The moas varied in size from the height of a turkey, to 3.7 metres high. Unfortunately, this made them easy targets, and they became extinct due to over-hunting by about 1500. As a result of this, the Māori switched back to agriculture.

The vocabulary coverage of the original text is _____%





② Suggest lexical changes in the text to make it easier for the students to read. Type or copy this text into LancsLex and check vocabulary coverage again. The vocabulary coverage of the modified text is _____%

③ Compare your solutions with your neighbour.



Task 17. Write a short paragraph (three or four sentences) appropriate for *intermediate learners of English*. The paragraph can be on any topic of your choice (e.g. sport, free time, university etc.) Copy or type the paragraph into LancsLex, check the vocabulary structure of your text and answer the following questions:

- What is the basic vocabulary coverage of your text?
- Are there any words that do not occur on the *new-GSL*?



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education





6 CQPWEB AND THE BNC

You will be using an online corpus search environment, **CQPweb** (<https://cqpweb.lancs.ac.uk/>) to access and analyse the BNC 1994 and the Spoken BNC 2014 (Love et al. 2017). **You need to create your CQPweb account, which can be done for free.** The ESRC Centre for Corpus Approaches to Social Science in Collaboration with Cambridge University Press have compiled a new corpus, the Spoken BNC2014, which consists of real-life, informal, spoken interactions between speakers of British English from across the United Kingdom. The corpus is available via Lancaster CQPweb.

The Spoken BNC2014 can be used, for example, to:

- Find frequencies of words, phrases and grammatical structures in current British speech.
- Find regional differences in spoken British English.
- Find sociolinguistic differences between different groups of speakers (based on gender, age, class, education...).

6.1 Image: BNC 1994 and BNC 2014



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education





T Task 18. Create CQPweb account and register for the use of the Spoken BNC 2014.

1. Sign up for an account on CQPweb: <https://cqpweb.lancs.ac.uk/>
2. Get access to the Spoken BNC2014: <http://corpora.lancs.ac.uk/bnc2014>

T Task 19. Searching & interpreting frequency information in CQPweb. Use the British National Corpus XML edition to find *the frequencies of the items below*. Tip: In CQPweb, you can use different ways of searching – you can search for a **word form** (a word or phrase exactly as you type it in), for a **lemma** using { }, or you can search for a **specific word class** using a tag such as ‘A’ for adjectives.

- a) lovely
 - b) research (used as a noun) SEARCH TIP: use a lemma search {research/N}
 - c) research (used as a verb) SEARCH TIP: use a lemma search {research/V}
- Compare the findings with your neighbour. Discuss the frequencies – which of the words is most frequent? Were there any problems with the search and finding the right forms (parts of speech) of the words?





CQPweb tip:
Type in words and expressions into the 'Standard query' search box. For advanced searches see 'Simple query language syntax' help file.

Menu
Corpus queries
Standard query
Restricted query
Word lookup
Frequency lists
Keywords

lovely



Task 20. Explore the Spoken BNC 2014. Go to CQPweb. From the main menu, select the Spoken BNC2014 corpus.

Let's think about colours. One of the claims that have been made about the differences between male and female language is that males use a more narrow range of colour terms than females. Look in the Spoken BNC2014 to see if this is true. Think of a list of colours terms and look at them in male speech and then do the same with female speech. What do you find? Are the collocates of individual colour terms the same in male and female speech?

From the main screen click the 'Restricted query' button on the left. This allows you to specify a query in the corpus with any of the restrictions available on the screen brought up by clicking this button. Check the box next to each feature you are interested in to place that restriction on the query you plan to run. To look for words spoken by males, scroll down and check the box for 'Male' under 'Gender'. Type your query in the search box, e.g. the word 'red'. This will give you the results for this word. Look at the top line on the screen for a summary of the results. For comparability, focus on the relative frequency (instances per million words).

Menu
Corpus queries
Standard query
Restricted query
Word lookup
Frequency lists
Keywords
Analyse corpus
Export corpus
Saved query data
Query history
Saved queries
Categorised queries
Upload a query
Create/edit subcorpora
Corpus info
View corpus metadata

Spoken BNC 2014

red

Query mode:

Number of hits per page:

Select the transcription conventions used
☐ Original
☐ Revised



EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education





To get back to the main screen, click ‘Go’ next to ‘New Query’ towards the top right of the screen. To get the results for females, check ‘Female’ rather than ‘Male’ under ‘Gender’.

You will need to run two queries for each colour term you think of – once with a ‘Male’ restriction, once with a ‘Female’ restriction. **So what are your conclusions about the use of colour terms by men and women in current spoken British English?**

T Task 21. Searching & comparing frequency information. Use the BNC 1994 and the Spoken BNC 2014 in CQPweb to find out whether the swearword ‘fuck’ as a verb and fuck as a noun occurs: i) more frequently in spoken than in written English and ii) occurs more often in spoken English in the 1990s or current spoken English. N.B. Use relative frequencies for comparison.

1. For RQ1, use the **British National Corpus** (XML edition). The British National Corpus, is a corpus of both spoken and written British English containing about 100 million words collected in the 1990s. In the ‘Restriction’ field choose ‘Written’ and then ‘Spoken’ to obtain the frequency for each mode of communication.

2. For RQ2, the data from the **British National Corpus (spoken part)** represent the language use in the 1990s. For the present-day use, go to **Spoken BNC2014**, a recently completed corpus that contains over 10 million words of informal spoken British English.

3. Additional RQs. If you have time, you can look up and compare the frequency of other words.

Word	BNC – Written part	BNC – Spoken part	Spoken BNC2014
fuck (as a noun)			
fuck (as a verb)			
Your own word			
Your own word			

- Compare the findings with your neighbour. Discuss what factors may affect the frequency of the occurrence of ‘fuck’ in each mode of communication and in the two time periods.





GLOSSARY OF CORPUS TERMINOLOGY

Absolute (or raw) frequency – The simple frequency with which a search term occurs in a corpus or its part(s); a number of hits of a search term in a corpus.

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

Collocation graph is a visual display of the association between a node and its collocates.

Collocation network is a visual display of complex associations (collocations) in language and discourse. It consists of multiple inter-connected collocation graphs.

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

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

Concordance line – A single line in the KWIC display representing a node (search term) with the words before and after it (the right and left context).

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

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

Dispersion – is the spread of values of a variable (e.g. relative frequencies of a search term) in a dataset (corpus). Dispersion is measured statistically using metrics such as standard deviation (*SD*), coefficient of variation (*CV*), range, Juilland's *D*, *DP* etc.

Frequency – The number of times a search term occurs in the corpus. A distinction is made between absolute (absolute number of hits) and relative frequency (proportional frequency per *X* number of tokens).

Lemma – All inflected forms belonging to one stem; in #LancsBox by default, a combination of a headword and a grammatical category (e.g. go + VERB). 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.





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.

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 any combination of strings.

Relative (or normalized) frequency (RF) is calculated as the proportion of the absolute frequency of a word we are interested in divided by the total number of words (tokens) in the corpus. This number is usually multiplied by an appropriate basis for normalization (e.g. 10,000).

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

Type is a unique word form in a text or corpus.





REFERENCES AND RECOMMENDED READINGS

- Baker, P. (2016). The shapes of collocation. *International Journal of Corpus Linguistics*, 21(2), 139-164.
- Brezina, V. (2018 forth.) *Statistics for corpus linguistics: A practical guide*. Cambridge: Cambridge University Press.
- Brezina, V. (2018.) Statistical choices in corpus-based discourse analysis. In: Taylor, Ch. & Marchi, A. (eds.) *Corpus approaches to discourse: a critical review*. Routledge: London.
- Brezina, V. (2016). Collocation Networks. In: Baker, P. & Egbert, J. (eds.) *Triangulating Methodological Approaches in Corpus Linguistic Research*. Routledge: London.
- Brezina, V. & Gablasova, D. (2018). The corpus method. In: Culpeper, J, Kerswill, P., Wodak, R., McEnery, T. & Katamba, F. (eds). *English Language (2nd edition)*. Palgrave.
- Brezina, V., & Gablasova, D. (2017). How to Produce Vocabulary Lists? Issues of Definition, Selection and Pedagogical Aims. A Response to Gabriele Stein. *Applied Linguistics*, 38(5), 764-767.
- Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics*, 36(1), 1-22.
- Brezina, V., McEnery, T. & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2), 139-173.
- Brezina, V., & Meyerhoff, M. (2014). Significant or random. A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics*, 19(1), 1-28.
- Ellis, R. (2015). *Understanding Second Language Acquisition*. Oxford: Oxford University Press.
- Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, 67 (S1), 155–179.
- Gablasova, D., Brezina, V., & McEnery, T. (2017). Exploring learner language through corpora: comparing and interpreting corpus frequency information. *Language Learning*, 67 (S1), 130–154.
- Gablasova, D., Brezina, V., Mcenery, T., & Boyd, E. (2017). Epistemic stance in spoken L2 English: The effect of task and speaker style. *Applied Linguistics*, 38(5), 613-637.
- Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014. *International Journal of Corpus Linguistics*, 22(3), 319-344.
- McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

