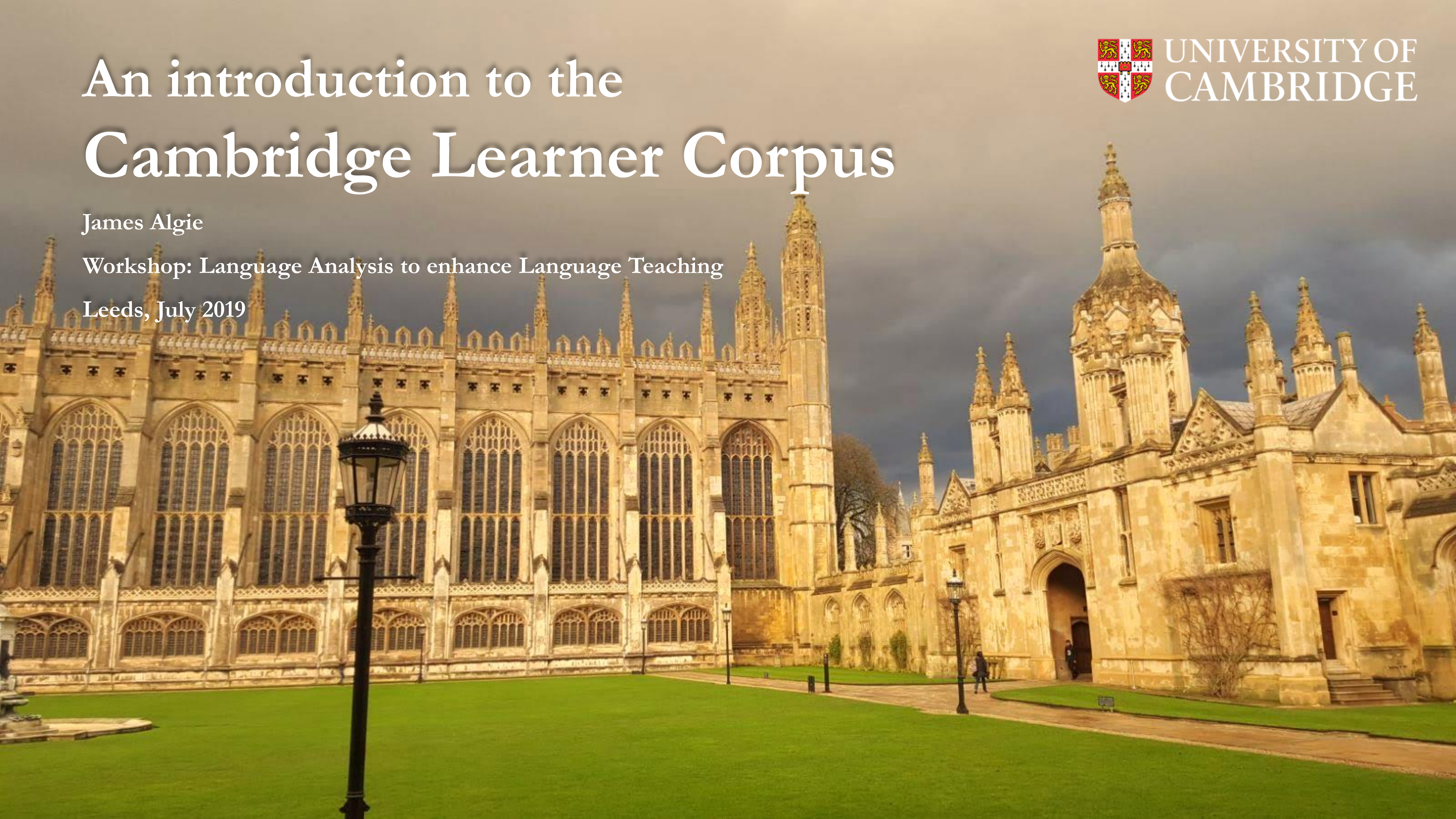


UNIVERSITY OF CAMBRIDGE

An introduction to the Cambridge Learner Corpus

James Algie

Workshop: Language Analysis to enhance Language Teaching

Leeds, July 2019

UNIVERSITY OF CAMBRIDGE

An introduction to the Cambridge Learner Corpus

- What is the Cambridge Learner Corpus?
- Where does the data come from?
- How can I access the corpus?
- What can I do with the corpus?
- What are the possible outcomes?

UNIVERSITY OF CAMBRIDGE

What is the Cambridge Learner Corpus?

- The Cambridge Learner Corpus (CLC) belongs to a wider range of corpora compiled and administered by Cambridge University Press and Cambridge Assessment English

- Cambridge English Corpus – includes Cambridge Reference Corpus (2bn words, expert users) and Cambridge Learner Corpus

CAMBRIDGE ENGLISH CORPUS
CAMBRIDGE QUALITY GUARANTEE

Where does the data come from?

UNIVERSITY OF CAMBRIDGE

- Written data produced by L2 English learners in language exams
- 55 million words produced by learners in 173 countries (and counting!)
- A1 to C2 learners – exam answers from various exams, including Cambridge English (all levels), CELS, IELTS, ESOL Skills for Life, Business and Legal
- Over half of the corpus (29 million words) is corrected and error-coded – a key feature and relatively rare for a corpus of this size
- Metadata included for each script - L1, nationality, age, exam level/performance/date, register/style/format – can be used to filter search results and as variables in statistical models

How can I access the corpus?

UNIVERSITY OF CAMBRIDGE

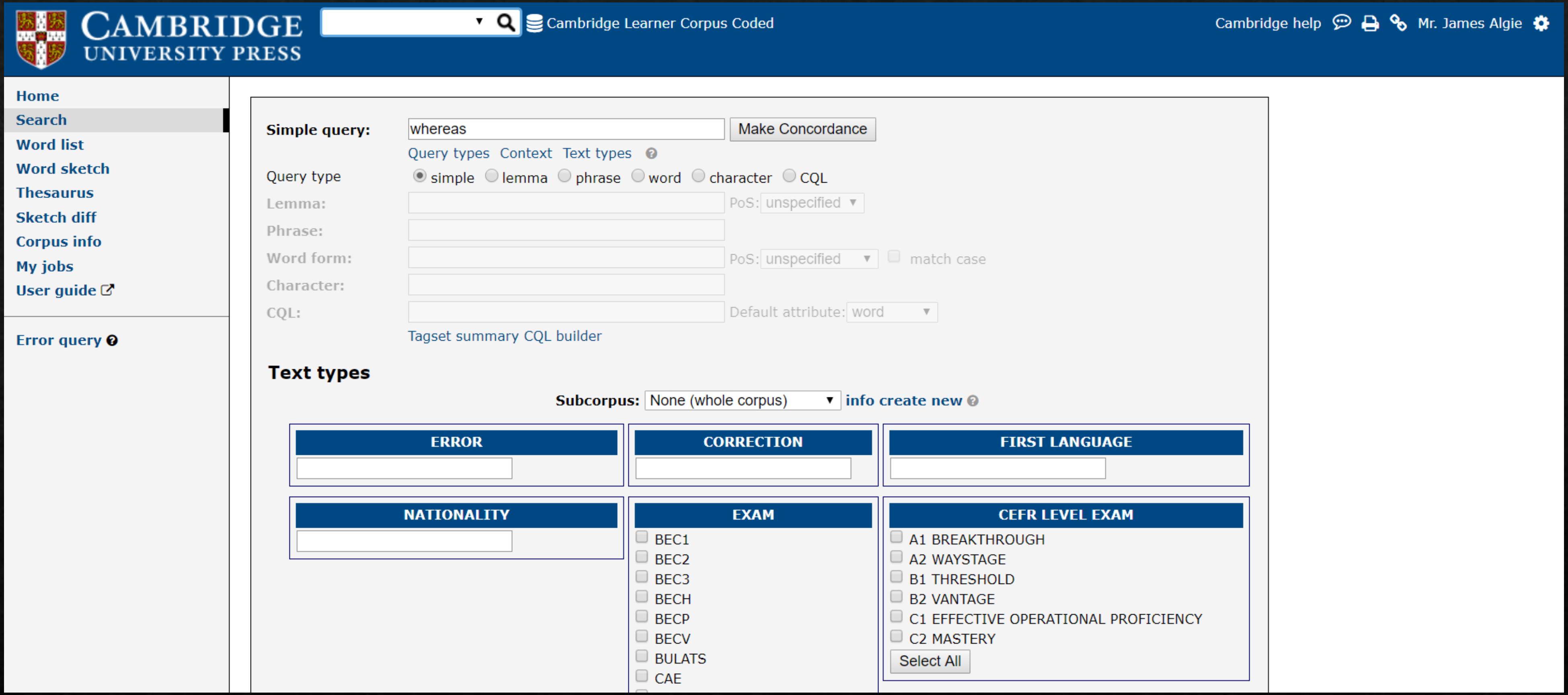
- Sketch Engine – provides several standard tools for corpus queries (word lists, word sketch, sketch diff, text types) as well as some enabled features for the CLC

- Usernames/passwords available for researchers and teachers on request: http://languageresearch.cambridge.org/academic-research-request-form

- Comprehensive user guides available for Sketch Engine, CLC and Corpus Query Language (CQL)

SKETCH ENGiNE

What can I do with the corpus?



Simple query: whereas

Query type ● simple ○ lemma ○ phrase ○ word ○ character ○ CQL

Lemma: PoS: unspecified

Phrase:

Word form: PoS: unspecified □ match case

Character:

CQL: Default attribute: word

Text types

Subcorpus: None (whole corpus)

ERROR

CORRECTION

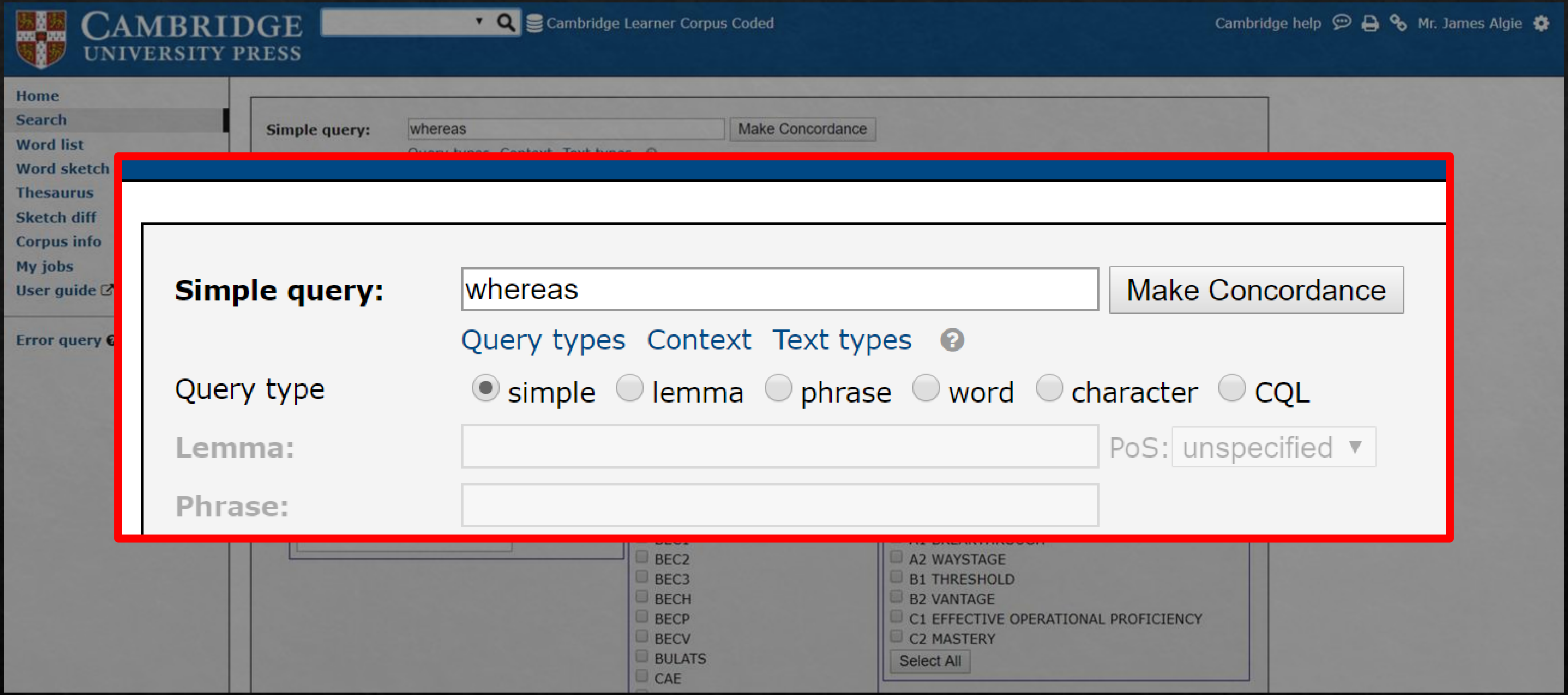
FIRST LANGUAGE

NATIONALITY

EXAM
- □ BEC1
- □ BEC2
- □ BEC3
- □ BECH
- □ BECP
- □ BECV
- □ BULATS
- □ CAE

CEFR LEVEL EXAM
- □ A1 BREAKTHROUGH
- □ A2 WAYSTAGE
- □ B1 THRESHOLD
- □ B2 VANTAGE
- □ C1 EFFECTIVE OPERATIONAL PROFICIENCY
- □ C2 MASTERY

What can I do with the corpus?



Simple query: whereas

Query type ● simple ○ lemma ○ phrase ○ word ○ character ○ CQL

Lemma:

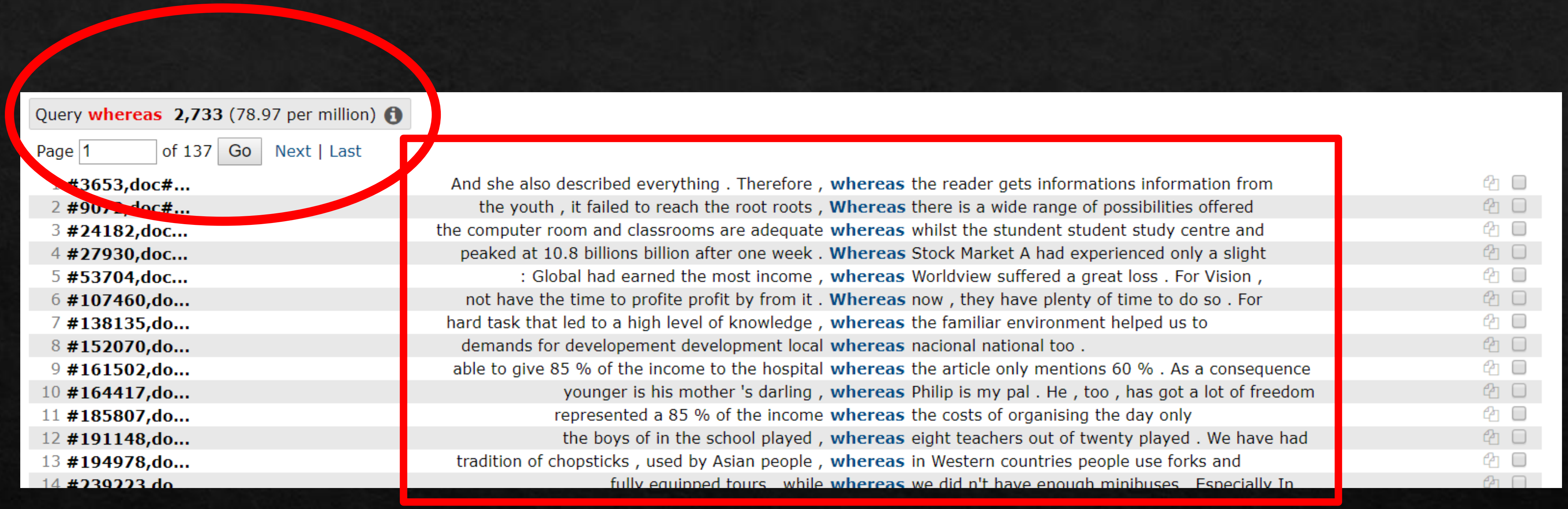
Phrase:

PoS: unspecified

Simple query

What can I do with the corpus?

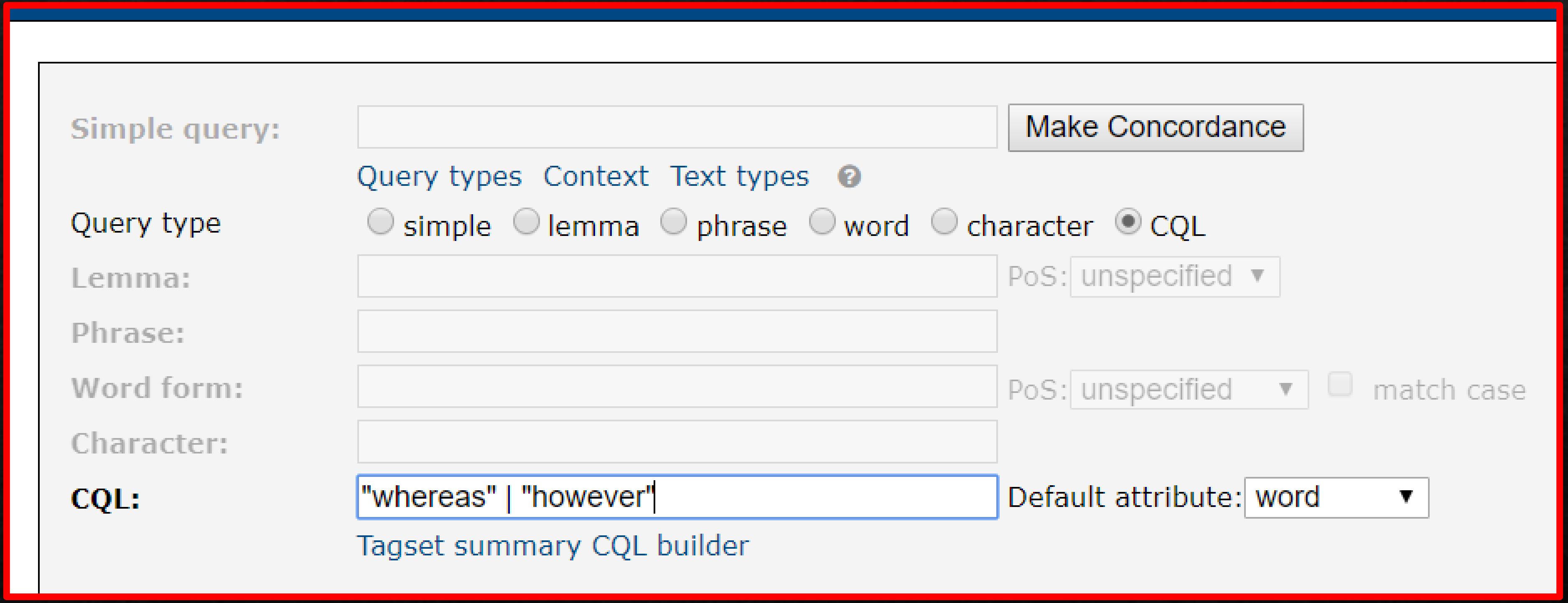
UNIVERSITY OF CAMBRIDGE



```
Query whereas 2,733 (78.97 per million)
Page 1 of 137 Go  Next | Last
 1 #3653,doc#...              And she also described everything . Therefore , whereas the reader gets informations information from
 2 #9072,doc#...                  the youth , it failed to reach the root roots , Whereas there is a wide range of possibilities offered
 3 #24182,doc...         the computer room and classrooms are adequate whereas whilst the stundent student study centre and
 4 #27930,doc...             peaked at 10.8 billions billion after one week . Whereas Stock Market A had experienced only a slight
 5 #53704,doc...                        : Global had earned the most income , whereas Worldview suffered a great loss . For Vision ,
 6 #107460,do...              not have the time to profite profit by from it . Whereas now , they have plenty of time to do so . For
 7 #138135,do...            hard task that led to a high level of knowledge , whereas the familiar environment helped us to
 8 #152070,do...           demands for developement development local whereas nacional national too .
 9 #161502,do...          able to give 85 % of the income to the hospital whereas the article only mentions 60 % . As a consequence
10 #164417,do...                        younger is his mother 's darling , whereas Philip is my pal . He , too , has got a lot of freedom
11 #185807,do...                       represented a 85 % of the income whereas the costs of organising the day only
12 #191148,do...                        the boys of in the school played , whereas eight teachers out of twenty played . We have had
13 #194978,do...             tradition of chopsticks , used by Asian people , whereas in Western countries people use forks and
14 #239223,do                            fully equipped tours , while whereas we did n't have enough minibuses . Especially In
```

Simple query

What can I do with the corpus?



Simple query: Make Concordance

Query types Context Text types

Query type ○ simple ○ lemma ○ phrase ○ word ○ character ● CQL

Lemma: PoS: unspecified ▼

Phrase:

Word form: PoS: unspecified ▼ □ match case

Character:

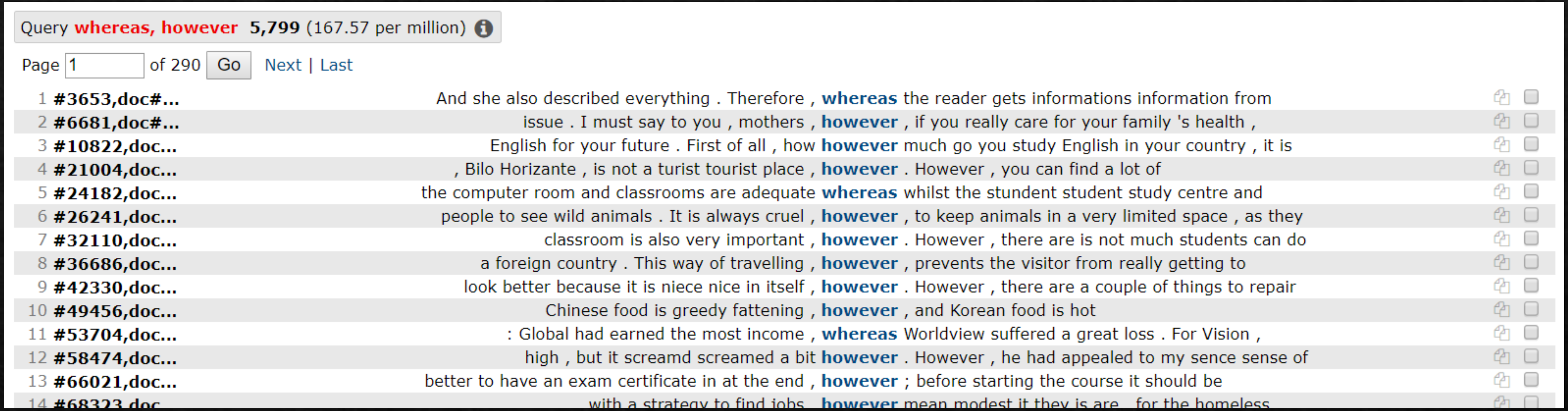
CQL: "whereas" | "however" Default attribute: word ▼

Tagset summary CQL builder

CQL search

What can I do with the corpus?

UNIVERSITY OF CAMBRIDGE



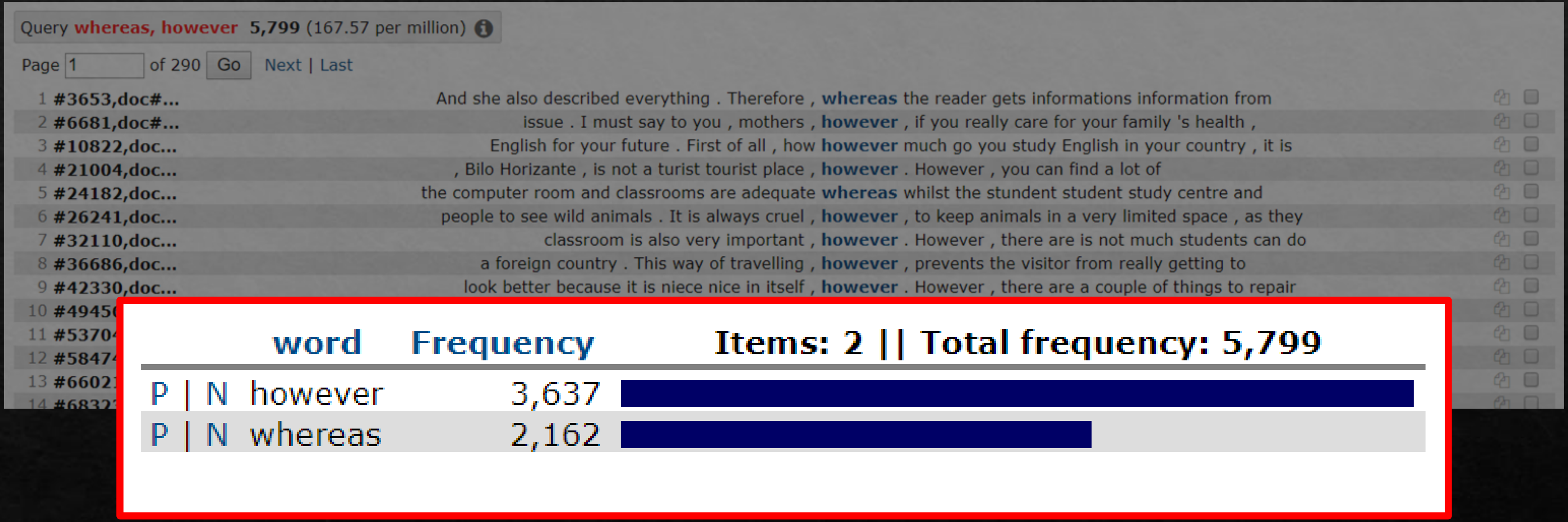
Query **whereas, however** **5,799** (167.57 per million)

Page 1 of 290 Go Next | Last

1	**#3653,doc#...**	And she also described everything . Therefore ,	**whereas**	the reader gets informations information from	□
2	**#6681,doc#...**	issue . I must say to you , mothers ,	**however**	, if you really care for your family 's health ,	□
3	**#10822,doc...**	English for your future . First of all , how	**however**	much go you study English in your country , it is	□
4	**#21004,doc...**	, Bilo Horizante , is not a turist tourist place ,	**however**	. However , you can find a lot of	□
5	**#24182,doc...**	the computer room and classrooms are adequate	**whereas**	whilst the stundent student study centre and	□
6	**#26241,doc...**	people to see wild animals . It is always cruel ,	**however**	, to keep animals in a very limited space , as they	□
7	**#32110,doc...**	classroom is also very important ,	**however**	. However , there are is not much students can do	□
8	**#36686,doc...**	a foreign country . This way of travelling ,	**however**	, prevents the visitor from really getting to	□
9	**#42330,doc...**	look better because it is niece nice in itself ,	**however**	. However , there are a couple of things to repair	□
10	**#49456,doc...**	Chinese food is greedy fattening ,	**however**	, and Korean food is hot	□
11	**#53704,doc...**	: Global had earned the most income ,	**whereas**	Worldview suffered a great loss . For Vision ,	□
12	**#58474,doc...**	high , but it screamd screamed a bit	**however**	. However , he had appealed to my sence sense of	□
13	**#66021,doc...**	better to have an exam certificate in at the end ,	**however**	; before starting the course it should be	□
14	**#68323,doc...**	with a strategy to find jobs ,	**however**	mean modest it they is are , for the homeless	□

CQL search

What can I do with the corpus?



Query **whereas, however** **5,799** (167.57 per million)

Page 1 of 290 Go Next | Last

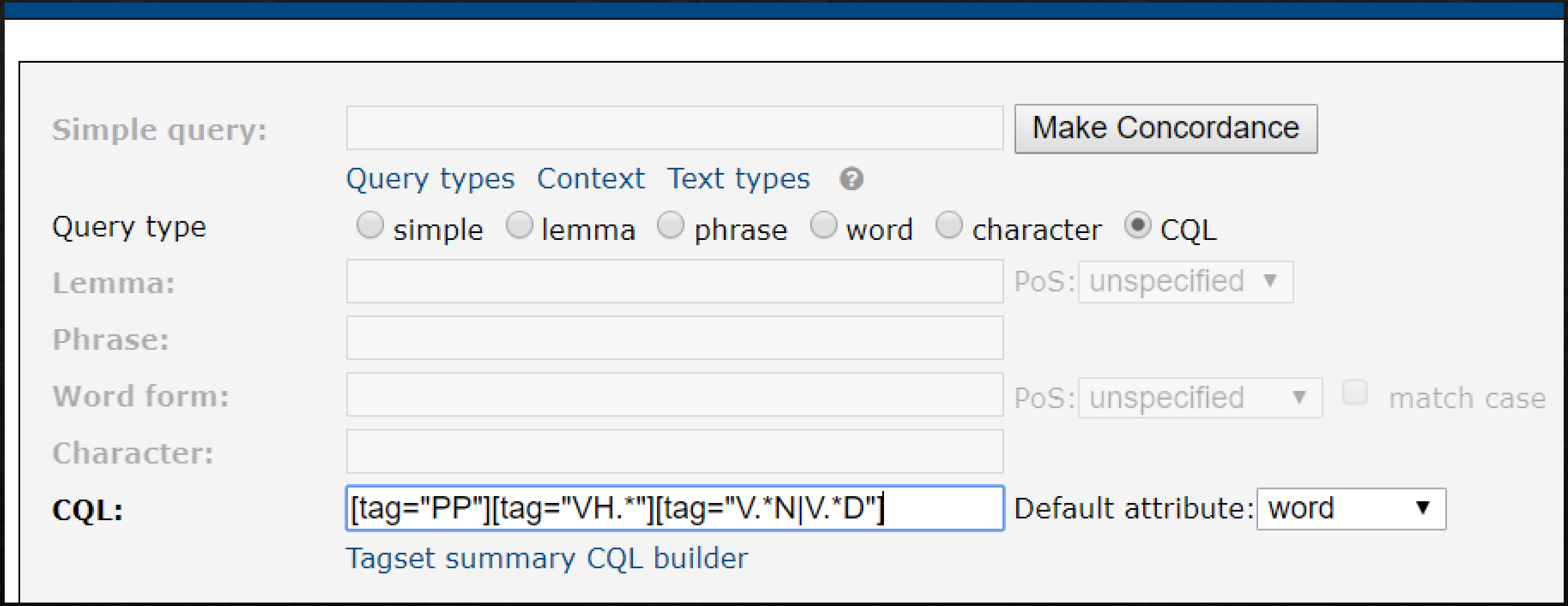
1	#3653,doc#...	And she also described everything . Therefore ,	**whereas**	the reader gets informations information from
2	#6681,doc#...	issue . I must say to you , mothers ,	**however**	, if you really care for your family 's health ,
3	#10822,doc...	English for your future . First of all , how	**however**	much go you study English in your country , it is
4	#21004,doc...	, Bilo Horizante , is not a turist tourist place ,	**however**	. However , you can find a lot of
5	#24182,doc...	the computer room and classrooms are adequate	**whereas**	whilst the stundent student study centre and
6	#26241,doc...	people to see wild animals . It is always cruel ,	**however**	, to keep animals in a very limited space , as they
7	#32110,doc...	classroom is also very important ,	**however**	. However , there are is not much students can do
8	#36686,doc...	a foreign country . This way of travelling ,	**however**	, prevents the visitor from really getting to
9	#42330,doc...	look better because it is niece nice in itself ,	**however**	. However , there are a couple of things to repair
10	#49456			
11	#53704			
12	#58474			
13	#66021			
14	#68323			

	word	Frequency	Items: 2 || Total frequency: 5,799
P | N	however	3,637	
P | N	whereas	2,162	

CQL search

What can I do with the corpus?

UNIVERSITY OF CAMBRIDGE



Simple query:

Query type ○ simple ○ lemma ○ phrase ○ word ○ character ● CQL

Lemma: PoS: unspecified

Phrase:

Word form: PoS: unspecified □ match case

Character:

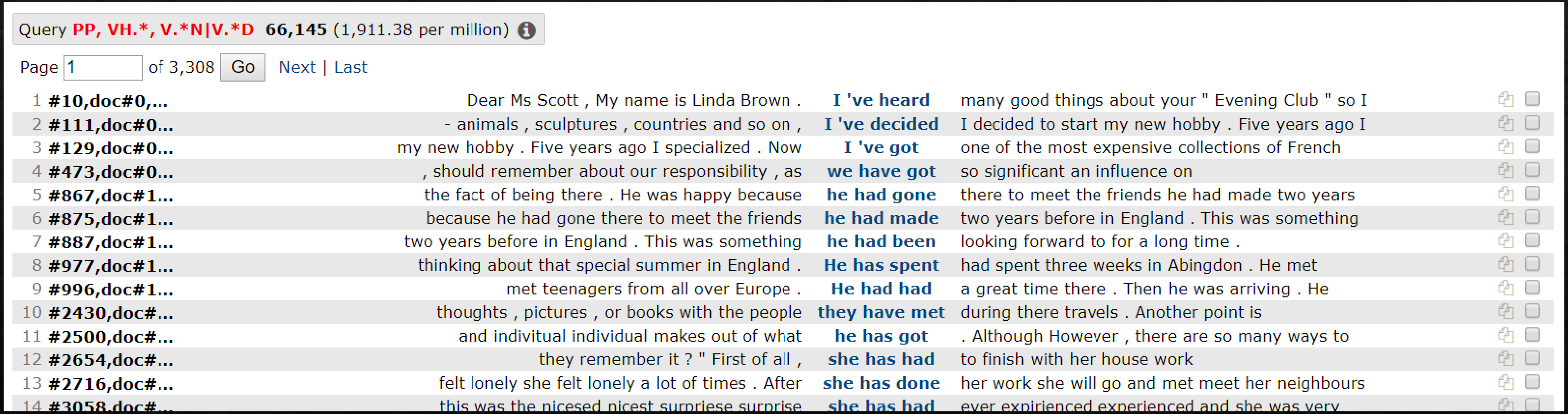
CQL: [tag="PP"][tag="VH.*"][tag="V.*N|V.*D"]

Default attribute: word

CQL search

What can I do with the corpus?

UNIVERSITY OF CAMBRIDGE



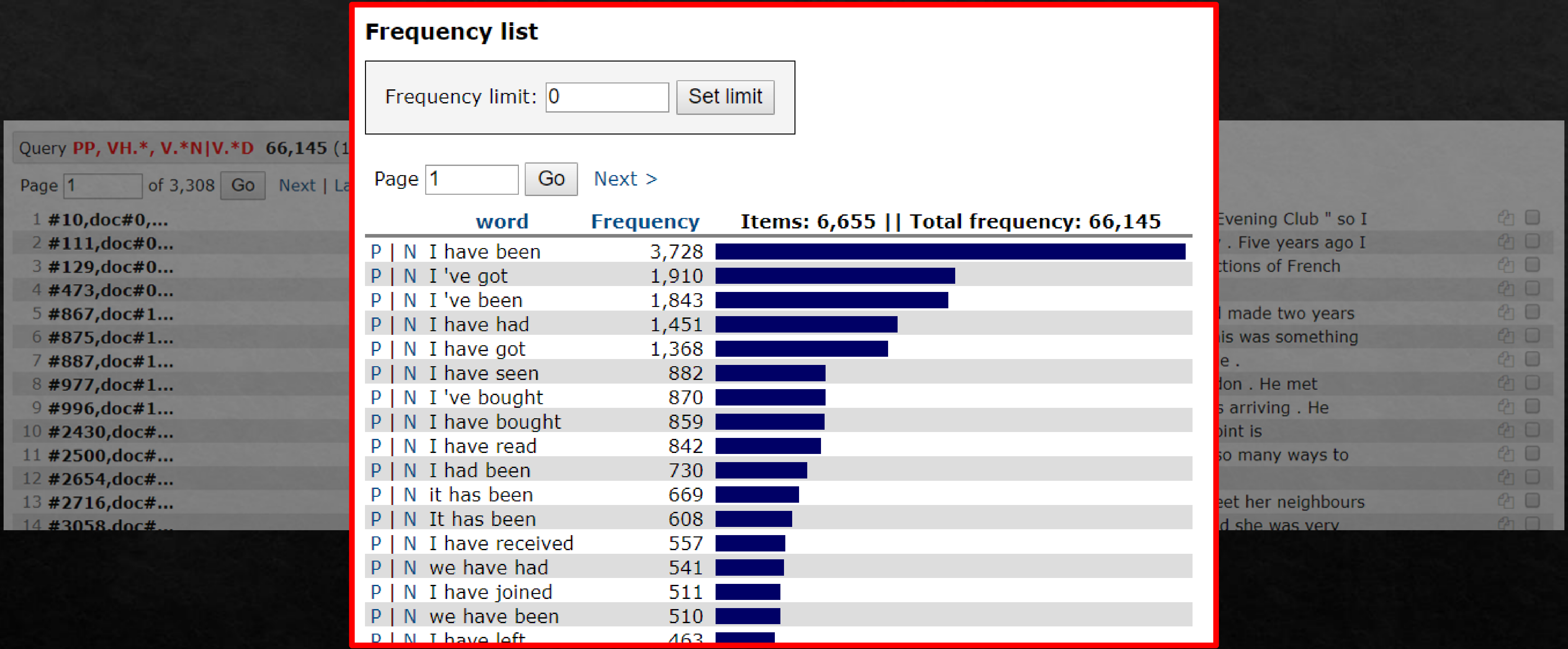
Query PP, VH.*, V.*N|V.*D 66,145 (1,911.38 per million)

Page 1 of 3,308 Go Next | Last

1	#10,doc#0,...	Dear Ms Scott , My name is Linda Brown .	I 've heard	many good things about your " Evening Club " so I
2	#111,doc#0...	- animals , sculptures , countries and so on ,	I 've decided	I decided to start my new hobby . Five years ago I
3	#129,doc#0...	my new hobby . Five years ago I specialized . Now	I 've got	one of the most expensive collections of French
4	#473,doc#0...	, should remember about our responsibility , as	we have got	so significant an influence on
5	#867,doc#1...	the fact of being there . He was happy because	he had gone	there to meet the friends he had made two years
6	#875,doc#1...	because he had gone there to meet the friends	he had made	two years before in England . This was something
7	#887,doc#1...	two years before in England . This was something	he had been	looking forward to for a long time .
8	#977,doc#1...	thinking about that special summer in England .	He has spent	had spent three weeks in Abingdon . He met
9	#996,doc#1...	met teenagers from all over Europe .	He had had	a great time there . Then he was arriving . He
10	#2430,doc#...	thoughts , pictures , or books with the people	they have met	during there travels . Another point is
11	#2500,doc#...	and indivitual individual makes out of what	he has got	. Although However , there are so many ways to
12	#2654,doc#...	they remember it ? " First of all ,	she has had	to finish with her house work
13	#2716,doc#...	felt lonely she felt lonely a lot of times . After	she has done	her work she will go and met meet her neighbours
14	#3058,doc#...	this was the nicesed nicest surpriese surprise	she has had	ever expirienced experienced and she was very

CQL search

What can I do with the corpus?



Query PP, VH.*, V.*N|V.*D 66,145 (1

Page 1 of 3,308 Go Next | La

1 #10,doc#0,...
2 #111,doc#0...
3 #129,doc#0...
4 #473,doc#0...
5 #867,doc#1...
6 #875,doc#1...
7 #887,doc#1...
8 #977,doc#1...
9 #996,doc#1...
10 #2430,doc#...
11 #2500,doc#...
12 #2654,doc#...
13 #2716,doc#...
14 #3058,doc#...

Evening Club " so I
. Five years ago I
tions of French

made two years
is was something
e .
don . He met
s arriving . He
int is
so many ways to

eet her neighbours
d she was very

Frequency list

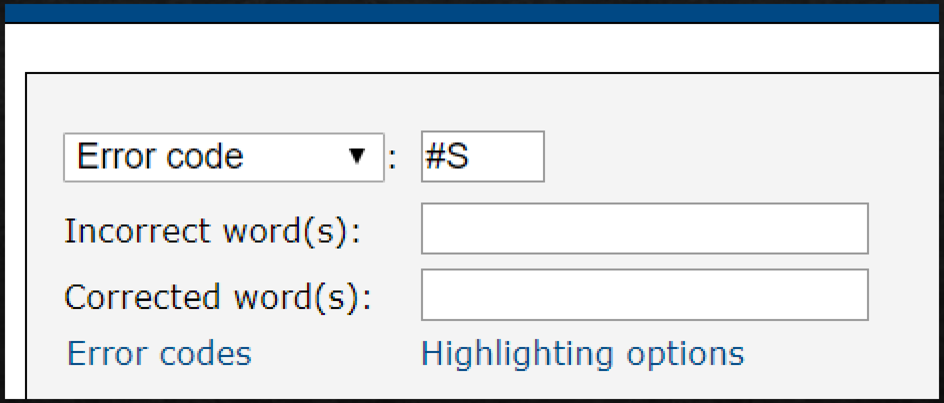
Frequency limit: 0 Set limit

Page 1 Go Next >

	word	Frequency	Items: 6,655 || Total frequency: 66,145
P | N	I have been	3,728	
P | N	I 've got	1,910	
P | N	I 've been	1,843	
P | N	I have had	1,451	
P | N	I have got	1,368	
P | N	I have seen	882	
P | N	I 've bought	870	
P | N	I have bought	859	
P | N	I have read	842	
P | N	I had been	730	
P | N	it has been	669	
P | N	It has been	608	
P | N	I have received	557	
P | N	we have had	541	
P | N	I have joined	511	
P | N	we have been	510	
P | N	I have left	463	

CQL search

What can I do with the corpus?



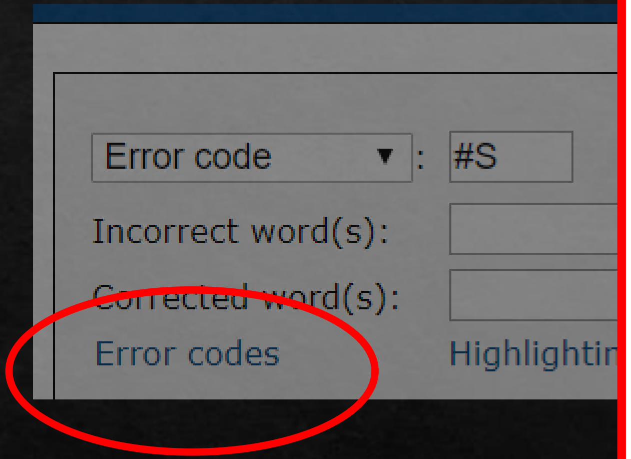
Error code ▼ : #S

Incorrect word(s):

Corrected word(s):

Search by error type

What can I do with the corpus?



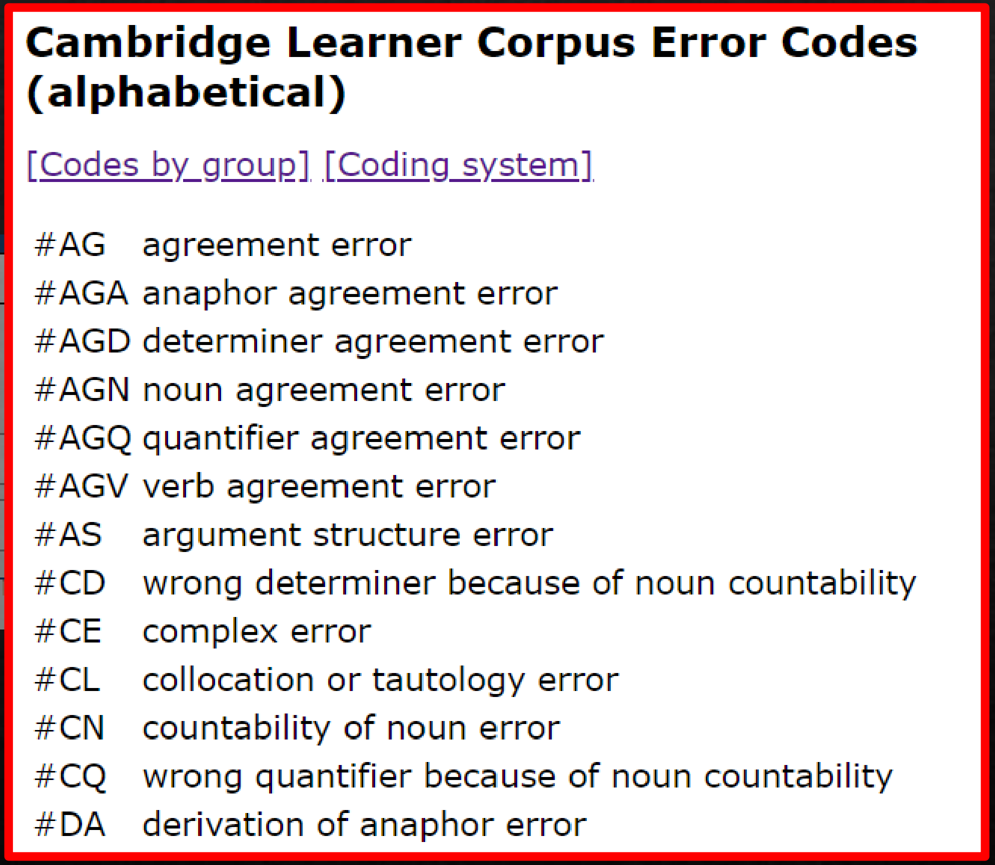
Error code ▼ : #S

Incorrect word(s):

Corrected word(s):

Error codes

Highlightin



Cambridge Learner Corpus Error Codes (alphabetical)

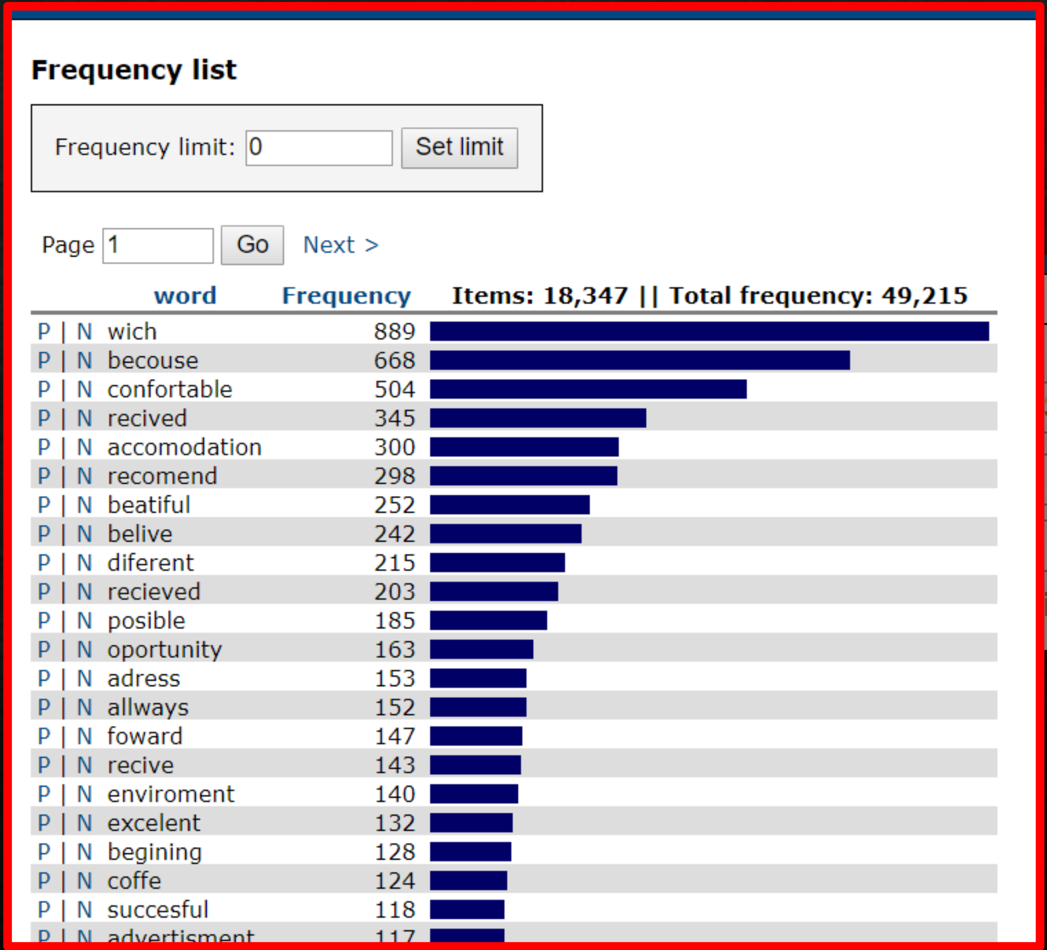
[Codes by group] [Coding system]

#AG	agreement error
#AGA	anaphor agreement error
#AGD	determiner agreement error
#AGN	noun agreement error
#AGQ	quantifier agreement error
#AGV	verb agreement error
#AS	argument structure error
#CD	wrong determiner because of noun countability
#CE	complex error
#CL	collocation or tautology error
#CN	countability of noun error
#CQ	wrong quantifier because of noun countability
#DA	derivation of anaphor error

Search by error type

What can I do with the corpus?

UNIVERSITY OF CAMBRIDGE



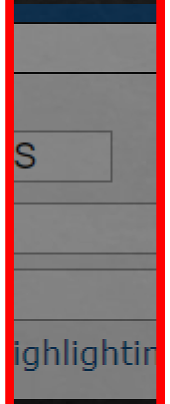
Frequency list

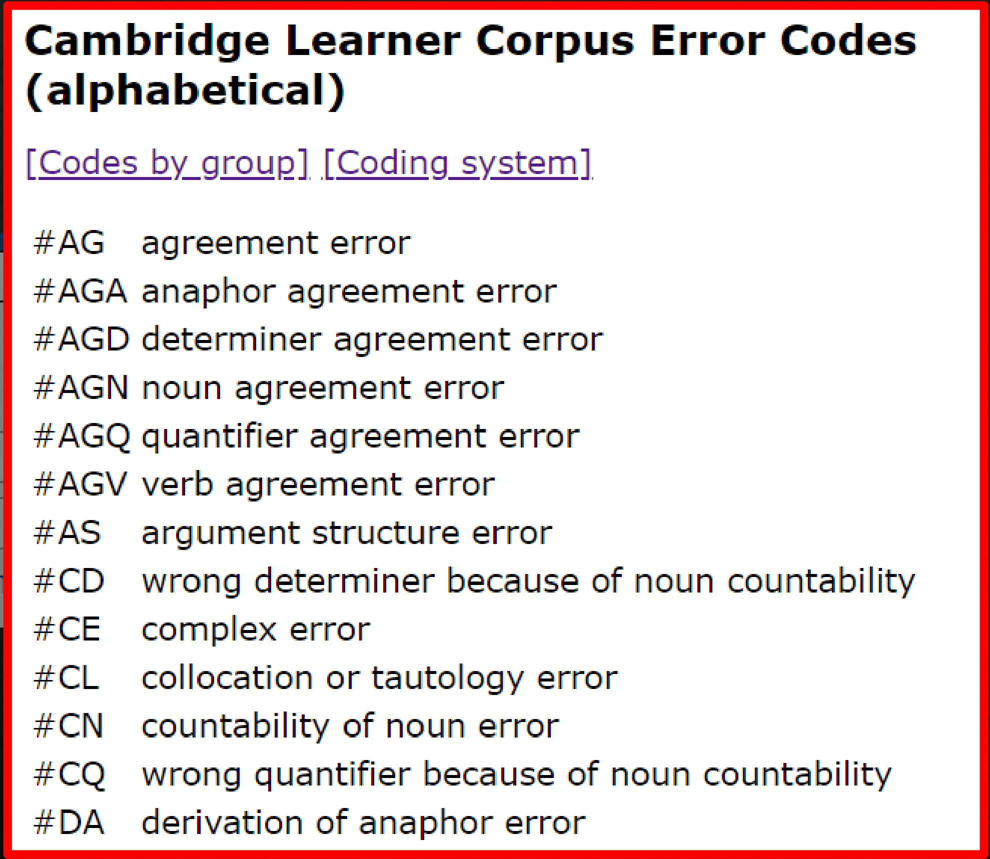
Frequency limit: 0 Set limit

Page 1 Go Next >

Items: 18,347 || Total frequency: 49,215

		word	Frequency
P	N	wich	889
P	N	becouse	668
P	N	confortable	504
P	N	recived	345
P	N	accomodation	300
P	N	recomend	298
P	N	beatiful	252
P	N	belive	242
P	N	diferent	215
P	N	recieved	203
P	N	posible	185
P	N	oportunity	163
P	N	adress	153
P	N	allways	152
P	N	foward	147
P	N	recive	143
P	N	enviroment	140
P	N	excelent	132
P	N	begining	128
P	N	coffe	124
P	N	succesful	118
P	N	advertisment	117





Cambridge Learner Corpus Error Codes (alphabetical)

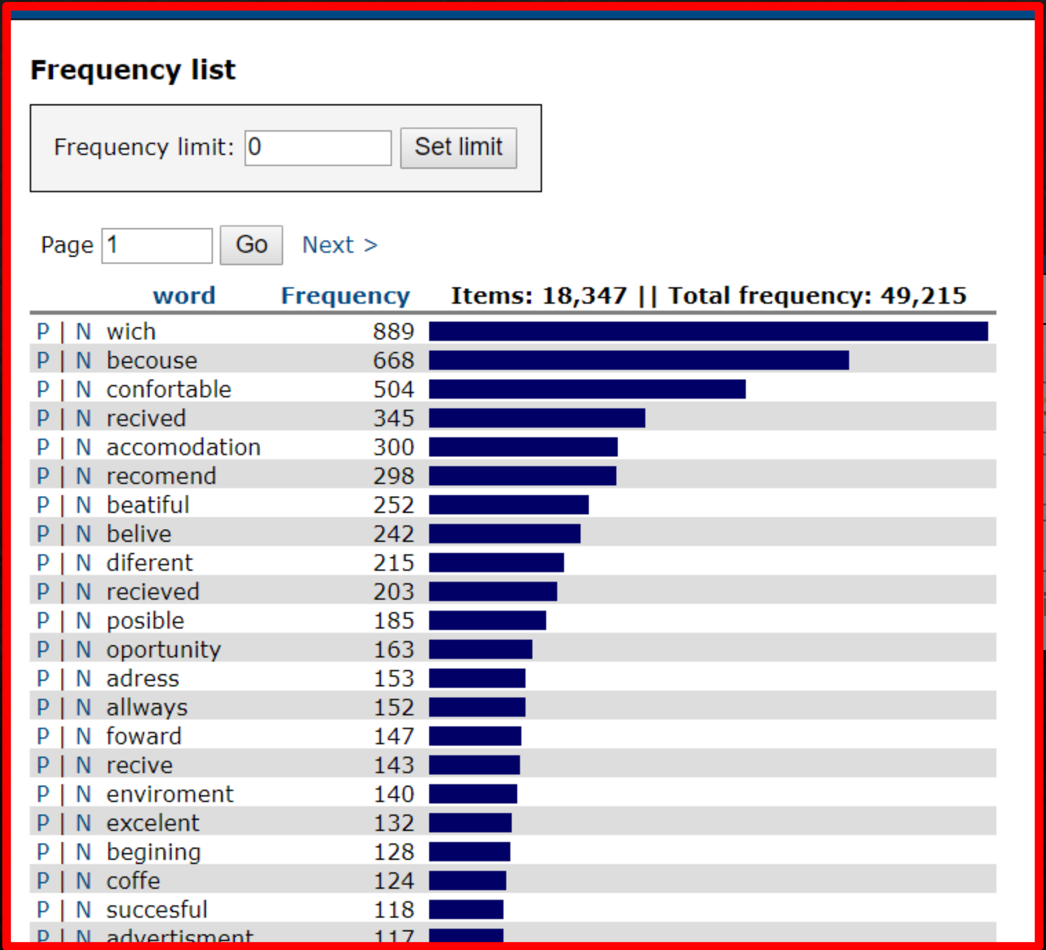
[Codes by group] [Coding system]

#AG	agreement error
#AGA	anaphor agreement error
#AGD	determiner agreement error
#AGN	noun agreement error
#AGQ	quantifier agreement error
#AGV	verb agreement error
#AS	argument structure error
#CD	wrong determiner because of noun countability
#CE	complex error
#CL	collocation or tautology error
#CN	countability of noun error
#CQ	wrong quantifier because of noun countability
#DA	derivation of anaphor error

#S

Search by error type

What can I do with the corpus?



Frequency list

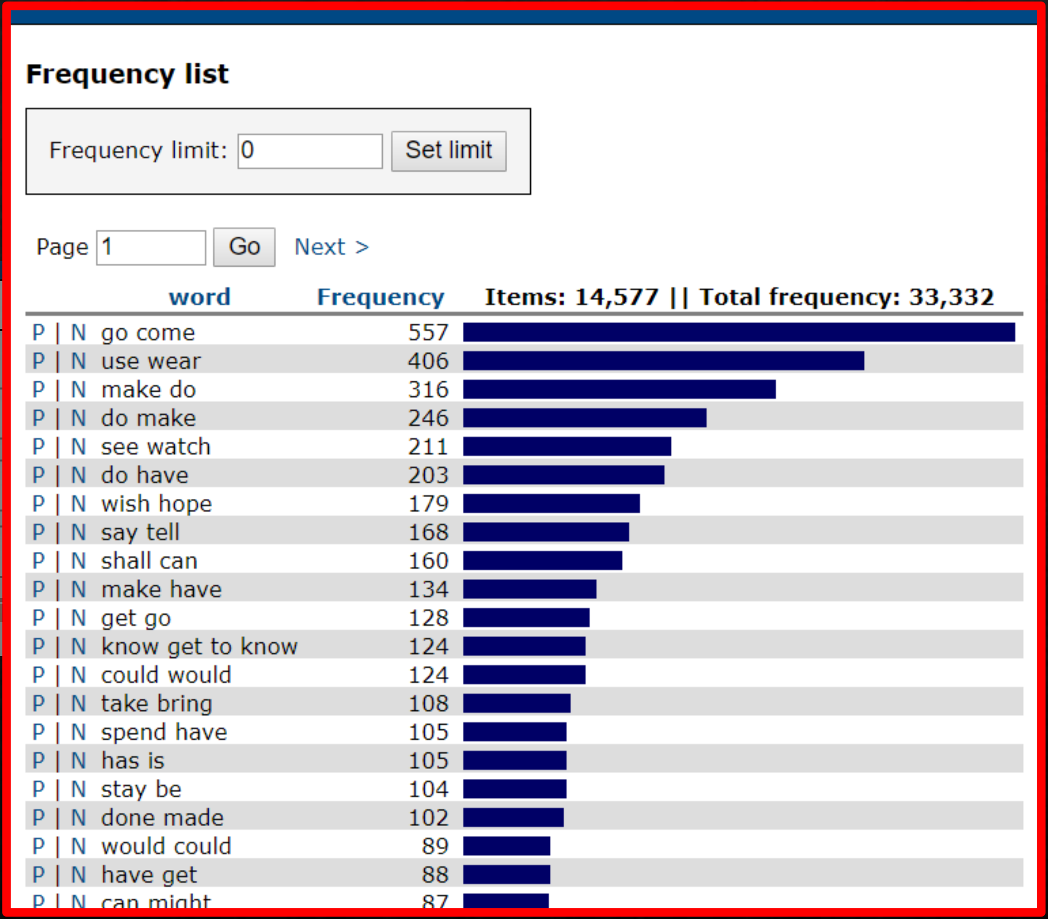
Frequency limit: 0 Set limit

Page 1 Go Next >

Items: 18,347 || Total frequency: 49,215

		word	Frequency
P	N	wich	889
P	N	becouse	668
P	N	confortable	504
P	N	recived	345
P	N	accomodation	300
P	N	recomend	298
P	N	beatiful	252
P	N	belive	242
P	N	diferent	215
P	N	recieved	203
P	N	posible	185
P	N	oportunity	163
P	N	adress	153
P	N	allways	152
P	N	foward	147
P	N	recive	143
P	N	enviroment	140
P	N	excelent	132
P	N	begining	128
P	N	coffe	124
P	N	succesful	118
P	N	advertisment	117

#S



Frequency list

Frequency limit: 0 Set limit

Page 1 Go Next >

Items: 14,577 || Total frequency: 33,332

		word	Frequency
P	N	go come	557
P	N	use wear	406
P	N	make do	316
P	N	do make	246
P	N	see watch	211
P	N	do have	203
P	N	wish hope	179
P	N	say tell	168
P	N	shall can	160
P	N	make have	134
P	N	get go	128
P	N	know get to know	124
P	N	could would	124
P	N	take bring	108
P	N	spend have	105
P	N	has is	105
P	N	stay be	104
P	N	done made	102
P	N	would could	89
P	N	have get	88
P	N	can might	87

#RV

Search by error type

What are the possible outcomes?

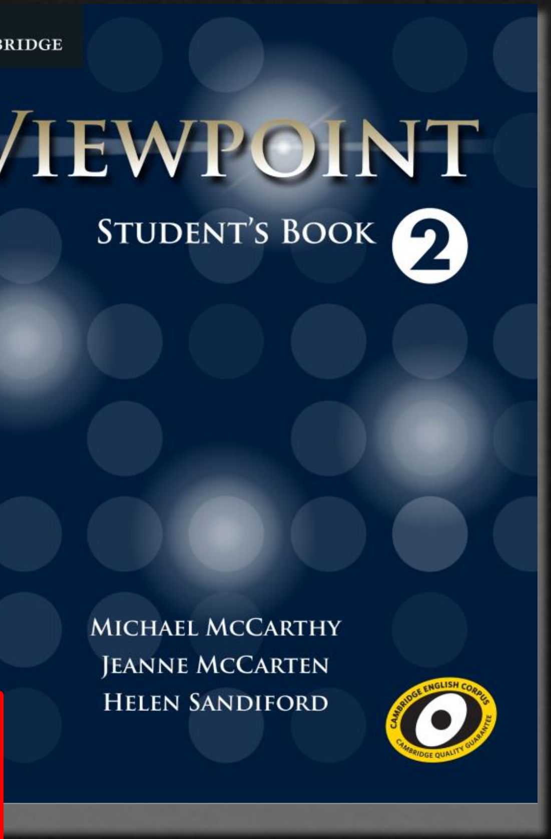
Teachers:

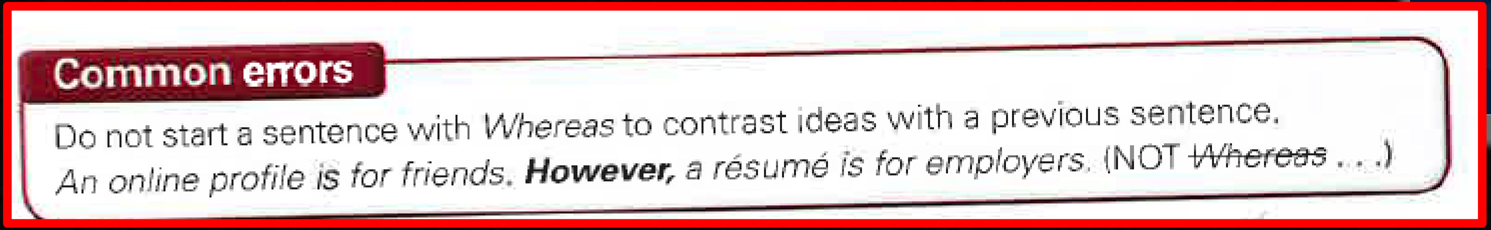
- Observe frequency of particular error types (among a particular demographic)
- Use contextualised examples as a pedagogical tool
  - *What are the most frequent spelling errors among B2 Polish learners?*
  - *Which mass nouns cause most problems for advanced learners?*

What are the possible outcomes?

Development of learning materials:

- Use common patterns and errors to inform textbooks and online resources
- Highlight problematic areas at appropriate stages of learning





Common errors

Do not start a sentence with *Whereas* to contrast ideas with a previous sentence.
An online profile is for friends. ***However,*** *a résumé is for employers.* (NOT ~~*Whereas*~~ . . .)

What are the possible outcomes?

Researchers:

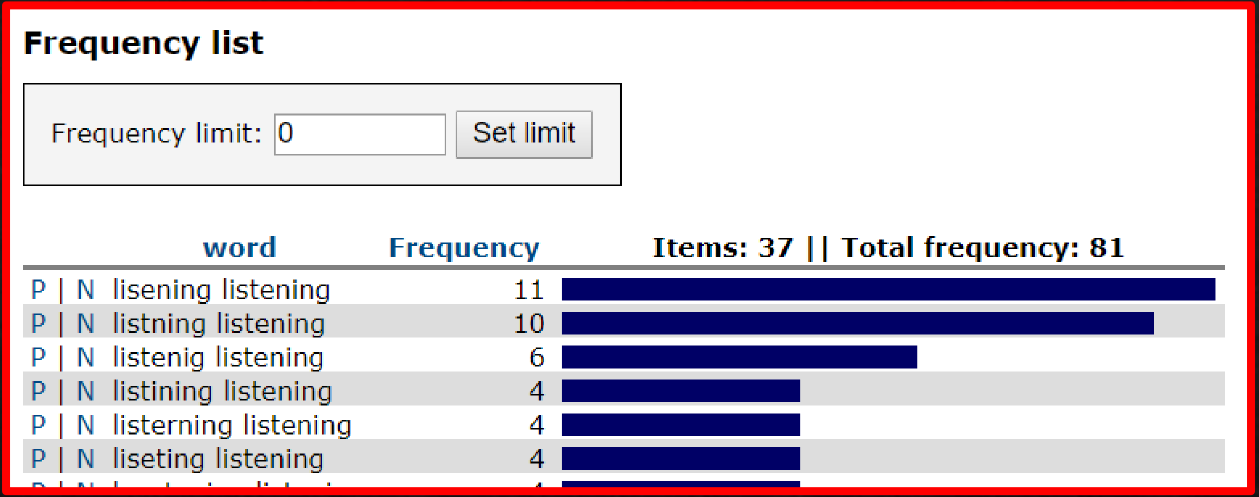
- Use CLC as an exploratory tool to inform research questions or as a method to address specific research questions
- Use rich metadata to incorporate multiple variables into statistical analyses
- Original and corrected/coded corpora can be combined to conduct 'Labovian' studies of learner behaviour
    - *Examining the relationship between L2 proficiency and variety and quantitative usage of adverbs* (Buttery & Caines 2012)
    - *Identifying criterial features to improve CEFR level descriptors* (Hawkins & Filipović 2012)
    - *Investigating L1 influence on the acquisition order of English grammatical morphemes* (Murakami & Alexopoulou 2016)

Thank you for ~~lisening~~ listening!

UNIVERSITY OF CAMBRIDGE

For access:
http://languageresearch.cambridge.org/academic-research-request-form

James Algie
ja600@cam.ac.uk
algie.j@cambridgeenglish.org



Frequency list

Frequency limit: 0 Set limit

	word	Frequency	Items: 37 || Total frequency: 81
P | N	lisening listening	11	
P | N	listning listening	10	
P | N	listenig listening	6	
P | N	listining listening	4	
P | N	listerning listening	4	
P | N	liseting listening	4	

Buttery, P. and Caines, A., 2012. Normalising frequency counts to account for ‘opportunity of use’ in learner corpora. *Developmental and Crosslinguistic Perspectives in Learner Corpus Research.* Amsterdam: John Benjamins Publishing Company, 4, pp.187-204.

Hawkins, J.A. & Filipović, L., 2012. *English Profile Studies 1: Criterial Features in L2 English.* Cambridge: Cambridge University Press.

Murakami, A. and Alexopoulou, T., 2016. L1 influence on the acquisition order of English grammatical morphemes: A learner corpus study. *Studies in Second Language Acquisition*, 38, pp.365– 401.