innovation in corpus linguistics.

#LancsBox

@Lancaster University

What's new in #LancsBox v. 6?

Version 6.0

- Create corpora from the web feature.
- Faster data loading and searches.
- Using space bar to shift between graph views in GraphColl.

Version 5.1

- Much more powerful searches throughout the tool.
- Uploading wordlists both in .txt and .csv.

Version 5.0

- Wizard: a new tool for automatic creation of research reports
- Shared collocates table (GraphColl)
- Clearer progress indication throughout the tool
- Faster searches

1 Create corpora

#LancsBox allows you to create a new corpus based on URLs provided. It downloads multiple webpages and extracts text from them.

1. In the corpora tab, left-click on 'Corpus' under 'Create'.

2. This will open a window where you can enter the corpus name and URLs, and specify extraction criteria.

1

Create Corpus	×
Corpus name	Linguistics
Enter URLs (1 per line)
https://en.wikipedia.or	g/wiki/Linguistics
Maximum pages	100 - Maximum levels 2 - Allow external links retrieval
Heuristics settings	● None ○ Page number ○ Date
Value matching URL	0
Maximum value	0 🔹
Patterns	<u>^.*(\? &)page={page}{&.*\$ \$), ^.*(\? &)pg={pa</u>
	Create corpus

3. By default, #LancsBox extracts 100 webpages at two levels of embedding.

4. Left-click on 'Create corpus' to start the extraction process.

5. When the process is finished, #LancsBox will notify you. You can then open the log with details about the websites identified and extracted and view the folder with the data (txt).

📕 Info	rmation	×
1	Finished!	
	The corpus "Linguistics" has been downloaded into #LancsB	OX.
	<u>Read log</u> <u>Preview corpus files</u>	
	ОК	

6. Click 'Ok' and then 'Import' to load data into #LancsBox directly.

Note: #LancsBox allows heuristic searches for websites that include multiple pages or dates e.g. <u>https://www.mumsnet.com/Talk/am_i_being_unreasonable/4170918-to-even-consider-giving-husband-second-chance?pg=1</u>

Switch on the heuristics setting and specify i) the heuristic type (page number or date) ii) the value matching the URL (1 in the example above) and iii) the maximum value. -1 Indicates unlimited number of pages until all pages are downloaded, while a date in the future indicates an unlimited number of dates until all dates are exhausted.

2 New search capabilities of #LancsBox

In #LancsBox, you can:

- \checkmark search for a word or phrase
- ✓ search for number ranges, e.g. >1930&<=1945
- ✓ use * wildcards, e.g. new*
- ✓ use regular expressions, e.g. /new c.*/i
- ✓ use 'smart searches', e.g. PASSIVE, NOUN
- ✓ use CQL, e.g. [word="go" & pos="V.*"]

<u>Complex searches.</u> From #LancsBox v. 5.1, extended search functionality is available. #LancsBox automatically identifies different search conventions and performs the desired search automatically. The user just needs to type the pattern in the main search box (KWIC, Whelk, GraphColl) or Wizard search box.

KWIC GraphColl Whelk Corpora KWIC X Wizard X ADVERB ADJECTIVE cat Search Search Occurrences Texts KWIC GraphColl Whelk Words Ngrams Corpora KWIC X Vizard X Image: Corpora to use Image: Corpora to use <td< th=""><th>#Lancsbox v J.1</th><th></th><th></th><th></th><th></th><th></th><th></th></td<>	#Lancsbox v J.1						
Corpora KWIC X Wizard X ADVERBADJECTIVE cat Search Search Occurrences Texts F #LancsBox v 5.1 Corpora KWIC X Wizard X Corpora kuse Vizard X Corpora to use Vizard X Corpora to use Vizard X Corpora to use Corpora to use Folds to use Search terms Output Brown L-C-M-C GraphColl Settings Whelk Folds Whelk Folds Whelk Settings Trext Settings	KWIC	Gr	aphColl	W	nelk		
ADVERSADJECTIVE cat Search Search Occurrences Texts F #LancsBox v 5.1 Corpora KWIC X Wizard X Corpora to use Frown L-C-M-C From Search terms Folde Whelk Words Settings Output Folde Whelk Words Settings Folde Folde Folde Folde	Corpora KWIC 🗙	Wizard 🗙					
Search Occurrences Texts Image: teanceBox v 5.1 Image: teanceBox v 5.1 Image: teanceBox v 5.1 Image: teanceBox v 5.1 Image: teanceBox v 5.1 Image: teanceBox v 5.1 Corpora to use GraphColl Whelk Words Ngrams Image: teanceBox v 5.1 Corpora to use Tools to use Tools to use Search terms Output Brown Image: teanchold teanphColl Settings Image: teanphColl Folde Image: LCCM-C Image: teanphColl Settings Image: teanphColl Folde Image: teanphColl Settings Image: teanphColl Settings Image: teanphColl Folde Image: teanphColl Settings Image: teanphColl Settings Image: teanphColl Folde Image: teanphColl Image: teanphColl Settings Image: teanphColl		cat		S	earch		
KWIC GraphColl Whelk Words Ngrams Corpora kwic X Wizard X Corpora to use Tools to use Search terms Output Brown CaphColl Settings NOLIN L-C-M-C GraphColl Settings Norans Whelk Words Settings Folde Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of the settings Image: Comparison of t	Search	Occurren	ces	Texts			
KWIC GraphColl Whelk Words Ngrams Corpora KWIC X Wizard X Corpora to use Search terms Outpute the settings Brown KWIC Settings L-C-M-C KWIC Settings Whelk Words Settings Whelk Words Settings Image: Corporal to use Settings							
KWIC GraphColl Whelk Words Ngrams Corpora to use Witard X Corpora to use Tools to use Search terms Outpu Brown CC.M-C GraphColl Settings INOUN UVhelk Words Settings Folde Words Settings Ngrams Folde	#LancsBox v 5.1						
Corpora k WK X Wizard X Corpora to use Corporato Corpora to use Corporato Corporat	KWIC	GraphColl	Whelk		Words	Ngrams	ſ
Corpora to use Tools to use Search terms Output Brown KWIC Settings INOUN C-M-C GraphColl Settings Nords Whelk Words Settings Ngrams Settings Text Settings	Corpora KWIC 🗙 Wiza	rd 🗙					
Brown C KWIC Settings NOUN Title C-C-M-C GraphColl Settings Folds Whelk Words Settings Text Settings	Corpora to use		Tools to use		Search terms		Outpu
GraphColl Settings Folds	Brown L-C-M-C		KWIC	Settings	PASSIVE		Title
Wheik Words Settings Ngrams Settings Text Settings			GraphColl	Settings			Folde
Words Settings Ngrams Settings Text Settings			Whelk				
□ Ngrams Settings □ Text Settings			Words	Settings			
Text Settings			Ngrams	Settings			
			Text	Settings			

The following conventions are available in complex searches.

- a) Multiple smart searches can be used in the same query; smart searches can also be combined with simple searches.
- 1) ADVERB ADJECTIVE
- 2) PRONOUN PASSIVE
- 3) ADVERB ADJECTIVE NOUN was

The following smart searches are available for English:

! , . ? ADJECTIVE ADVERB ΒE BOOSTER COLLECTIVE NOUN COMPARATIVE COMPLEX NOUN PHRASE CONDITIONAL CONNECTOR CONTRACTION DEGREE ADVERB DETERMINER DO DOWNTONER EXISTENTIAL THERE GERUND HAVE INFINITIVE HYPHENATED WORD **INDEFINITE PRONOUN** INFINITIVE INTERJECTION LINKING ADVERB LONG WORD MODAL NEGATION NOMINALIZATION NOUN NUMBER PARTICLE PASSIVE PAST TENSE PAST PARTICIPLE PERFECT INFINITIVE PHRASAL VERB PLACE ADVERB PREPOSITIONAL PHRASE PRESENT TENSE PRONOUN

PROPER NOUN REFLEXIVE PRONOUN REPETITION SHORT WORD SPLIT INFINITIVE SUPERLATIVE SWEARWORDS TIME ADVERB VERB

- b) The OR operator can be used in simple searches to indicate alternatives; it can be combined with parentheses to indicate which words belong together as shown in 3) and 4)
- 1) cat OR dog
- 2) car OR dog OR mouse
- 3) my (cat OR dog)
- 4) (my cat) OR (my dog)

Note: It is not possible to use the OR operator for combining expressions of different length – with a different number of words, e.g. (my cat) OR dog

c) The NOT operator can be used in simple searches to negate a search term (meaning 'anything but X'); it can be combined with parentheses to indicate which words belong together as shown in 3), 4) and 5).

NOT my
 NOT my friend
 NOT (my friend)
 NOT (a good) idea
 NOT (a good or bad) idea NOT me

d) #LancsBox also supports CQL (Corpus Query Language). It can be used for defining complex searches at different levels of annotation (1-4) or their combinations. All queries in CQL inside double quotes are interpreted as case insensitive regular expressions; for case sensitivity double equals sign (==) is required, e.g. 5).

CQL allows searching at the following levels of annotation: i) word, ii) headword (hw, lemma), iii) pos and iv) tag. While i)-iii) are supplied automatically for languages with full grammatical support, iv) represents an optional level of a user-defined tag. For example, a single item can be defined in CQL as

[word="goes" & headword="go" & pos="V.* "]

This is interpreted as a form of the word *goes* with the headword *go* and part-of speech tag V.* (verb). Note that the ampersand (&) is used to separate different levels of annotation inside square brackets. If a level of annotation is not specified, no restriction is applied at that level.

In CQL, square brackets [] separate slots in a phrase. 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 words without any restriction $([]{0,3})$ and followed by the past participle (V.N).

- 1) [word="cat"]
- 2) [headword="go"]
- 3) [pos="V.*"]
- 4) [tag="XX"]
- 5) [word== "Cat"]
- 6) [word="go" & headword="go" & pos="N.* "]
- 7) [headword="go" & pos="V.*"] [word="to"]
- 8) [headword="very" & pos="R.*"]{2} [pos="J.*"]
- 9)

3 The Wizard tool

The Wizard tool combines the power of all tools in #LancsBox, searches corpora and produces research reports for print (docx) and web (htlm).

It can be used, for example, to:

- Carry out simple or complex research.
- Produce a draft report.
- Download all relevant data.



The report produced by Wizard follows the structure of an academic research report – please see the example below.

Created by #LancsBox Wizard

Comparison of British and American English

1 Introduction

This research report was automatically produced by #LancsBox (Brezina et al. 2020), a corpus analysis tool developed at Lancaster University. It uses cutting-edge technology and statistical sophistication (Brezina 2018) to analyze and visualize corpus data. For more information and tips on research report writing see the <u>Research Report Guide</u>.

2 Method

2.1 Data

The study analyzed the following corpora:

Table 1. Corpora used

Name	Language	Texts	Tokens	Additional
D	E l'- h	45	4 94 4 9 5 4	T 40.000
Brown	English	15	1,014,361	Types: 49,686
				Lemmas: 44,622
L-O-B	English	15	1,007,677	Types: 48,349
				Lemmas: 43,920

In the study, 2 corpora were used of the total size of 2,022,038 running words (tokens) in 30 texts. A full description of the corpora is available in <u>data\tsv\corpora</u>.

2.2 Procedure

#LancsBox (Brezina et al. 2020) software package was employed to analyse the data. The following tool from the package was used: KWIC. The KWIC tool generates a list of all instances of a search term in a corpus in the form of a concordance. The following search terms were used: "new", "old" and "some".

3 Results