What’s new in #LancsBox v. 5?

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 New search capabilities of #LancsBox

In #LancsBox, you can:
- 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.*"]

Complex searches. 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.
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:

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

1) NOT my
2) NOT my friend
3) NOT (my friend)
4) NOT (a good) idea
5) NOT (a good or bad) idea


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).
2 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.
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 Research Report Guide.

2 Method

2.1 Data

The study analyzed the following corpora:

<table>
<thead>
<tr>
<th>Name</th>
<th>Language</th>
<th>Texts</th>
<th>Tokens</th>
<th>Additional information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>English</td>
<td>15</td>
<td>1,014,351</td>
<td>Types: 45,686</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lemmas: 44,622</td>
</tr>
<tr>
<td>L-O-B</td>
<td>English</td>
<td>15</td>
<td>1,007,677</td>
<td>Types: 48,349</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lemmas: 43,320</td>
</tr>
</tbody>
</table>

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 datasets\corpora.

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